



Statistical Policy
Working Paper 20

Seminar on Quality of Federal Data

Part 1 of 3

Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

March 1991

**MEMBERS OF THE FEDERAL COMMITTEE ON
STATISTICAL METHODOLOGY**

(February 1991)

**Maria E. Gonzalez, Chair
Office of Management and Budget**

**Yvonne M. Bishop
Energy Information
Administration**

**Warren L. Buckler
Social Security Administration**

**Charles E. Caudill
National Agricultural
Statistics Service**

**Cynthia Z.F. Clark
National Agricultural
Statistics Service**

**Zahava D. Doering
Smithsonian Institution**

**Robert M. Groves
Bureau of the Census**

**Roger A. Herriot
National Center for
Education Statistics**

**C. Terry Ireland
National Computer Security
Center**

**Charles D. Jones
Bureau of the Census**

**Daniel Kasprzyk
Bureau of the Census**

**Daniel Melnick
National Science Foundation**

**Robert P. Parker
Bureau of Economic Analysis**

**David A. Pierce
Federal Reserve Board**

**Thomas J. Plewes
Bureau of Labor Statistics**

**Wesley L. Schaible
Bureau of Labor Statistics**

**Fritz J. Scheuren
Internal Revenue Service**

**Monroe G. Sirken
National Center for
Health Statistics**

**Robert D. Tortora
Bureau of the Census**

PREFACE

In 1975, the Office of Management and Budget (OMB) organized the Federal Committee on Statistical Methodology. Comprised of individuals selected by OMB for their expertise and interest in statistical methods, the committee has during the past 15 years determined areas that merit investigation and discussion, and overseen the work of subcommittees organized to study particular issues. Since 1978, 19 Statistical Policy Working Papers have been published under the auspices of the Committee.

On May 23-24, 1990, the Council of Professional Associations on Federal Statistics (COPAFS) hosted a "Seminar on the Quality of Federal Data." Developed to capitalize on work undertaken during the past dozen years by the Federal Committee on Statistical Methodology and its subcommittees, the seminar focused on a variety of topics that have been explored thus far in the Statistical Policy Working Paper series. The subjects covered at the seminar included:

- Survey Quality Profiles
- Paradigm Shifts Using Administrative Records
- Survey Coverage Evaluation
- Telephone Data Collection
- Data Editing
- Computer Assisted Statistical Surveys
- Quality in Business Surveys
- Cognitive Laboratories
- Employer Reporting Unit Match Study
- Approaches to Developing Questionnaires
- Statistical Disclosure-Avoidance
- Federal Longitudinal Surveys

Each of these topics was presented in a two-hour session that featured formal papers and discussion, followed by informal dialogue among all speakers and attendees.

Statistical Policy Working Paper 20, published in three parts, presents the proceedings of the "Seminar on the Quality of Federal Data." In addition to providing the papers and formal discussions from each of the twelve sessions, this working paper includes Robert M. Groves' keynote address, "Towards Quality in a Working Paper Series on Quality," and comments by Stephen E. Fienberg, Margaret E. Martin, and Hermann Habermann at the closing session, "Towards an Agenda for the Future."

We are indebted to all of our colleagues who assisted in organizing the seminar, and to the many individuals who not only presented papers and discussions but also prepared these materials for publication. A special thanks is due to Terry Ireland and his staff for their work in assembling this working paper.

Table of Contents

Wednesday, May 23, 1990

Part 1

KEYNOTE ADDRESS

TOWARDS QUALITY IN A WORKING PAPER SERIES ON QUALITY.	3
Robert M. Groves, The University of Michigan and U. S. Bureau of the Census	

Session 1 - SURVEY QUALITY PROFILES

THE SIPP QUALITY PROFILE.	19
Thomas B. Jabine, Statistical Consultant	
INITIAL REPORT ON THE QUALITY OF AGRICULTURAL SURVEY PROGRAM. .	29
George A. Hanuschak, National Agricultural Statistics Service	
DISCUSSION.	40
Barbara A. Bailer, American Statistical Association	
DISCUSSION.	46
Nancy A. Mathiowetz, U. S. Bureau of the Census	

Session 2 - PARADIGM SHIFTS USING ADMINISTRATIVE RECORDS

PARADIGM SHIFTS: ADMINISTRATIVE RECORDS AND CENSUS-TAKING. . .	53
Fritz Scheuren, Internal Revenue Service	
AN ADMINISTRATIVE RECORD PARADIGM: A CANADIAN EXPERIENCE . . .	66
John Leyes, Statistics Canada	

DISCUSSION.	77
Gerald Gates, U.S. Bureau of the Census	
DISCUSSION.	83
Edward J. Spar, Market Statistics	

Session 3 - SURVEY COVERAGE EVALUATION

CONTROL MEASUREMENT, AND IMPROVEMENT OF SURVEY COVERAGE	87
Gary M. Shapiro, U. S. Bureau of the Census; Raymond R. Bosecker, National Agricultural Statistics Service	
QUALITY OF SURVEY FRAMES.100
Judith T. Lessler, Research Triangle Institute	
DISCUSSION.108
Fritz Scheuren, Internal Revenue Service	
DISCUSSION.114
Joseph Waksberg, Westat, Inc.	

Session 4 - TELEPHONE DATA COLLECTION

QUALITY IMPROVEMENT IN TELEPHONE SURVEYS.123
Leyla Mohadjer, David Morganstein, Westat, Inc.	
COMPUTER ASSISTED SURVEY TECHNOLOGIES IN GOVERNMENT: AN OVERVIEW.137
Marc Tosiano, National Agricultural Statistics Service	
DISCUSSION.155
William L. Nicholls II, U. S. Bureau of the Census	
DISCUSSION.161
James T. Massey, National Center for Health Statistics	

Part 2

Session 5 - DATA EDITING

OVERVIEW OF DATA EDITING IN FEDERAL STATISTICAL AGENCIES. . .	.167
David A. Pierce, Federal Reserve Board	
EDITING SOFTWARE (An excerpt from Chapter IV of Working Paper 18).173
Mark Pierzchala, National Agricultural Statistics Service	
RESEARCH ON EDITING180
Yahia Ahmed, Internal Revenue Service	
DISCUSSION.184
Charles E. Caudill, National Agricultural Statistics Service	
DISCUSSION.186
Richard Bolstein, George Mason University	

Session 6 - COMPUTER ASSISTED STATISTICAL SURVEYS

OVERVIEW OF COMPUTER ASSISTED SURVEY INFORMATION COLLECTION .	.191
Richard L. Clayton, U. S. Bureau of Labor Statistics	
A COMPARISON BETWEEN CATI AND CAPI.197
Martin Baum, National Center for Health Statistics	
COMPUTER ASSISTED SELF INTERVIEWING202
Ralph Gillmann, Energy Information Administration	
COMPUTER ASSISTED SELF INTERVIEWING: RIGS AND PEDRO, TWO EXAMPLES205
Ann M. Ducca, Energy Information Administration	
DATA COLLECTION209
Cathy Mazur, National Agricultural Statistics Service	

DISCUSSION.212
Robert N. Tinari, U. S. Bureau of the Census	
DISCUSSION.216
David Morganstein, Westat, Inc.	

Thursday, May 24, 1990

Session 7 - QUALITY IN BUSINESS SURVEYS

IMPROVING ESTABLISHMENT SURVEYS AT THE BUREAU OF LABOR STATISTICS221
Brian MacDonald, Alan R. Tupek, U. S. Bureau of Labor Statistics	
A REVIEW OF NONSAMPLING ERRORS IN FEDERAL ESTABLISHMENT SURVEYS WITH SOME AGRIBUSINESS EXAMPLES232
Ron Fecso, National Agricultural Statistics Service	
DISCUSSION.243
David A. Binder, Statistics Canada	
DISCUSSION.247
Charles D. Cowan, Opinion Research Corporation	

Session 8 - COGNITIVE LABORATORIES

THE BUREAU OF LABOR STATISTICS' COLLECTION PROCEDURES RESEARCH LABORATORY: ACCOMPLISHMENTS AND FUTURE DIRECTIONS253
Cathryn S. Dipbo, Douglas Herrmann, U. S. Bureau of Labor Statistics	
THE ROLE OF A COGNITIVE LABORATORY IN A STATISTICAL AGENCY.268
Monroe G. Sirken, National Center for Health Statistics	
DISCUSSION.278
Elizabeth Martin, U. S. Bureau of the Census	
DISCUSSION.281
Murray Aborn, National Science Foundation (retired)	

Session 11 - STATISTICAL DISCLOSURE - AVOIDANCE

DISCLOSURE AVOIDANCE PRACTICES AT THE CENSUS BUREAU367
Brian Greenberg, U. S. Bureau of the Census	
THE MICRODATA RELEASE PROGRAM OF THE NATIONAL CENTER FOR HEALTH STATISTICS377
Robert H. Mugge, National Center for Health Statistics (retired)	
DISCUSSION.385
George T. Duncan, Carnegie Mellon University	

Session 12 - FEDERAL LONGITUDINAL SURVEYS

FEDERAL LONGITUDINAL SURVEYS.393
Daniel Kasprzyk, U. S. Bureau of the Census; Curtis Jacobs, U. S. Bureau of Labor Statistics	
THE ADVANTAGES AND DISADVANTAGES OF LONGITUDINAL SURVEYS. . .	.407
Robert W. Pearson, Social Science Research Council	
LONGITUDINAL ANALYSIS OF FEDERAL SURVEY DATA.425
Patricia Ruggles, Joint Economic Committee	
DISCUSSION.438
Michael Brick, Westat, Inc.	
DISCUSSION.447
Marilyn E. Manser, U. S. Bureau of Labor Statistics	

TOWARDS AN AGENDA FOR THE FUTURE

Stephen E. Fienberg, Carnegie Mellon University455
Margaret E. Martin.462
Hermann Habermann, Office of Management and Budget.465

Part 3

Session 9 - EMPLOYER REPORTING UNIT MATCH STUDY

INTERAGENCY AGREEMENTS FOR MICRODATA ACCESS:	
THE ERUMS EXPERIENCE291
Thomas B. Petska, Internal Revenue Service; Lois Alexander, Social Security Administration	
SAMPLE SELECTION AND MATCHING PROCEDURES USED IN ERUMS.301
John Pinkos, Kenneth LeVasseur, Marlene Einstein, U. S. Bureau of Labor Statistics; Joel Packman, Social Security Administration	
RESULTS, FINDINGS, AND RECOMMENDATIONS OF THE ERUMS PROJECT309
Vern Renshaw, Bureau of Economic Analysis; Tom Jabine, Statistical Consultant	
DISCUSSION.318
W. Joel Richardson, Charles A. Waite, U. S. Bureau of the Census	
DISCUSSION.324
Thomas J. Plewes, U. S. Bureau of Labor Statistics	

Session 10 - APPROACHES TO DEVELOPING QUESTIONNAIRES

TOOLS FOR USE IN DEVELOPING QUESTIONS AND TESTING	
QUESTIONNAIRES331
Theresa J. DeMaio, U. S. Bureau of the Census	
TECHNIQUES FOR EVALUATING THE QUESTIONNAIRE DRAFT340
Deborah H. Bercini, National Center for Health Statistics	
DESIGNING QUESTIONNAIRES FOR CATI IN A MIXED MODE	
ENVIRONMENT.349
Gemma Furno, U. S. Bureau of the Census	
DISCUSSION.360
Carol C. House, National Agricultural Statistics Service	

Part 1

Keynote Address

**TOWARDS QUALITY IN A WORKING PAPER
SERIES ON QUALITY**

TOWARDS QUALITY IN A WORKING PAPER SERIES ON QUALITY

Robert M. Groves
The University of Michigan and
U.S. Bureau of the Census

1. Introduction

Although this meeting has the title of the "Seminar on the Quality of Federal Data," its structure follows quite closely the topics covered in the multi-paper series of Statistical Policy Working Papers sponsored by the Office of Statistical Policy and Standards. There are as of this date, 19 Statistical Policy Working Papers written since the first in 1978. That is about 1.6 per year over the 12 years of the series (see Figure 1). They range over a wide terrain, involving issues of the topical focus of surveys to a set of methodological and statistical issues affecting survey quality.

I am unaware of the processes that led to my being asked to give the keynote address at this meeting. I must admit that I speak to you today as someone who has a very biased opinion about the OMB Statistical Policy Working Papers - I love almost all of them; I like the idea that they exist and only recently, because of my change of job sectors, have I appreciated their worth from another perspective. I have used them in graduate courses for students in survey methods (they are fine introductions to important design topics). I have used them in my research work (they are unique sources of documentation about what goes on in the Federal Statistical System). I recommend them to others calling for consulting assistance.

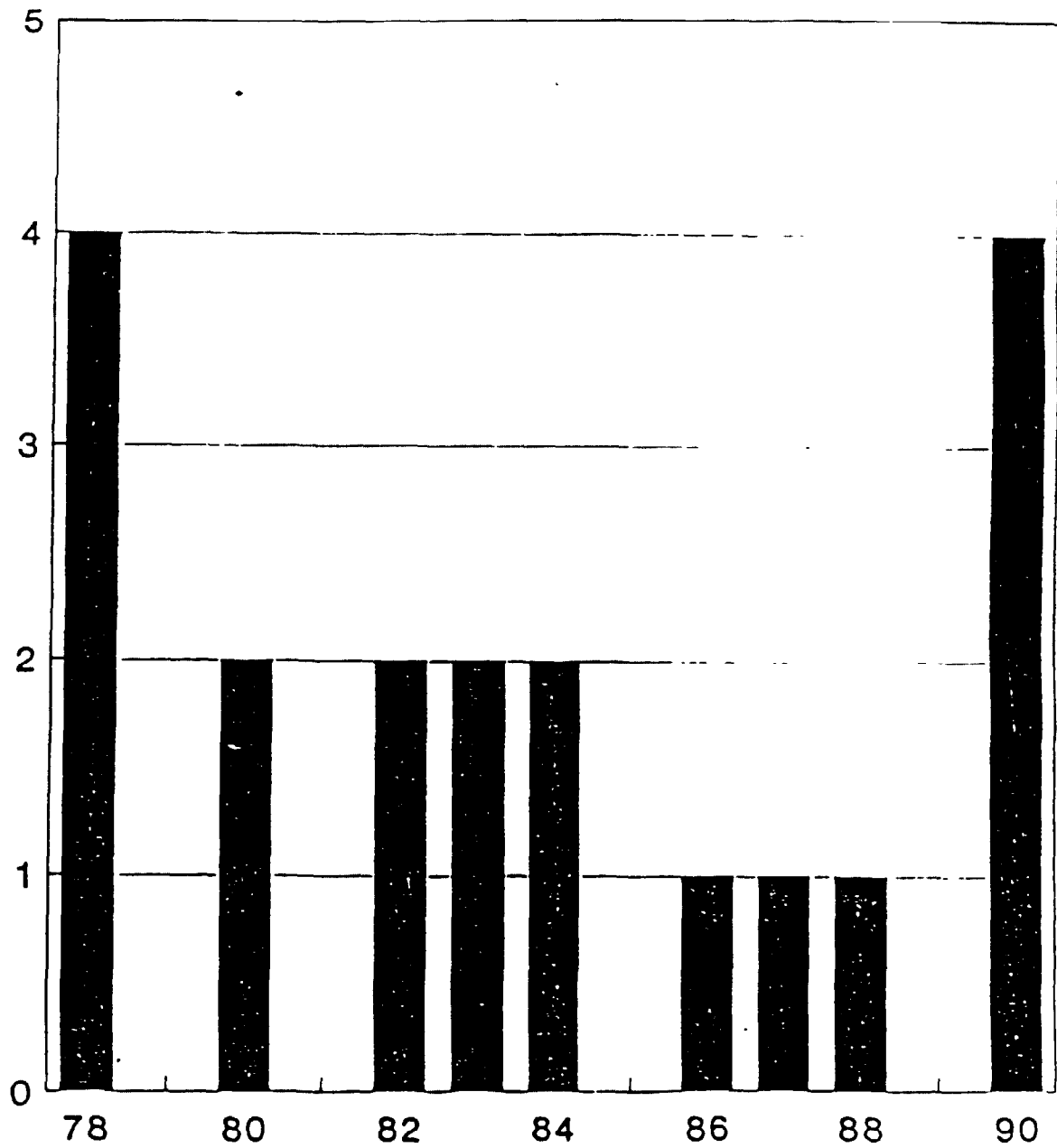
Although I speak as a friend, 45 minutes of praise from me wouldn't act to improve this series and runs the risk of "head inflation" for those who developed the papers. Instead, I want to be a constructive critic and will divide my remarks into several categories:

- a. alternative goals of the OMB series
- b. the need for a structure to their topics

I note that what follows are my personal views as a close observer from afar of the system and a rookie member of the system.

Figure 1

Year of Publication of OMB Working Papers



2. Alternative Perspectives on Goals of the Working Paper Series

2.1. OMB Series as Review of the State of Practice

Some of the papers in the series address a topic that spans many surveys of different populations (see Figure 2). The papers on coverage error and telephone data collection are examples of this. These kind of papers are compact summaries of the state of the art on a current issue facing all surveys. They often describe activities in both household surveys and those in economic surveys. Many times they end with case studies of different surveys across the Federal system and how they handle the particular issue at hand.

Figure 2

Alternative Perspectives on Goals of the Working Paper Series

1. OMB series as a review of the state of practice
2. OMB series as agency cross-fertilization
3. OMB series as a prod to new developments

These kind of papers are valuable to the extent that they have deep depth and wide breadth. By that I mean, they cover all the sources of data quality and cover them in sufficient depth that real learning is likely on the part of most readers.

Let me first speak of breadth of topics. I find it most simple to array the topics of the papers along the components of total survey error (see Figure 3). It is unfair for me to present this chart without some clarifying remarks about the missing cells. First, missingness does not imply absence of any treatment of the topics. Indeed, on sampling error, for example, many of the reports comment on the impact of design options on sampling variance. Second, this structure is only one which could be applied to classify the xx reports. Considering the label of this seminar "quality of Federal Data", however, I find it attractive to use it here.

Despite the weakness of any one classification scheme, let me point out what I believe are weaknesses with the current status of the series. There is a distinct bias toward the household survey domain to the detriment of the economic domain. There is one paper with the overarching title of "Quality in Establishment Surveys", but the fact that it alone exists underscores the problem. This is a reflection of the smaller literature in the methodology and evaluation of quality of economic surveys, but it is a status that

I hope will change in the future. Why? We have in the past too quickly assumed the following premises about economic survey measurement:

- a. establishment surveys are too diverse to yield themselves to common methodologies or standards.
- b. establishment surveys do not face questionnaire design issues like those of household surveys because the information gathered is factual in nature
- c. establishment surveys have nonresponse properties that do not resemble those of household surveys.

Each of these can be refuted with some observation of the various establishment surveys now ongoing. It is true that establishment populations have large variation in size; that their organizational structures are diverse; that their recordkeeping practices are not standardized; that the ideal respondent for different issues may vary across establishments. All of this is true, but should not lead to the extreme that there are no common problems either across different establishment surveys or between household and economic surveys.

As the Boskin report has observed, economic survey data needs improvement and the working paper series could be one vehicle of focusing attention on specific needs in this area.

The next most important omission, in my opinion, concerns the issue of nonresponse. I must admit here that the work of the National Academy of Sciences Panel on Missing and Incomplete Data offers a comprehensive review of current theory and practice. Conversely, the issue is vital to the unique inferential power of probability samples and therefore cannot receive too much attention. Even the most basic issues remain unresolved: relationships between response rates and nonresponse error; relationships between likelihood of coverage and likelihood of participation; cost/error evaluations of alternative methods of improving response rates. Mean square errors of survey estimators stem from thousands of individual decisions to cooperate with the survey request. It behooves us to devote more energy to this and the working paper series should do this.

Third, the interviewer has largely been ignored. It has been ignored despite that fact that many Federal surveys use interviewers to assist in the data collection, despite the fact that evaluative procedures desperately need review and reconceptualization, despite the fact that it is an area where both statistics and social science perspectives work. The attention to the interviewer is even more important given the likely future in which the traditional labor force of underemployed/overskilled part

time homemakers will decline and computer technologies are likely to transform the job.

Fourth, although large portions of data collection in the Federal Statistical System is by mail and self-administered questionnaire there is no focused treatment of the methodology in the series.

Fifth, a few comments specifically on error profiles. When I first read the CPS error profile 12 years ago, I had two reactions. I was attracted to the literary form -- a compilation of quality measures for the survey, combined with documentation of design features. I then felt and still believe that the structure of an error profile is a valuable way to document leading components of error in survey statistics (we should be grateful to Brooks and Bailer as the mothers (or midwives) of the invention). My second reaction came after digesting the full report. How little we as a community seemed to know about the error properties of the CPS, the largest ongoing and one of the most important ongoing Federal household surveys. Of the 80 pages of the report, for example, only about 25 are devoted to the data collection operations, a source of most of the errors in the process! That combination of reactions led me to the belief that I still have -- the error profile, in the hands of intelligent program directors, can act as an agenda setting document for quality improvement programs.

Finally, there are no serious treatments of costs of data collection - a topic I'll revisit in a few minutes.

Let me now turn to issues of depth. At their worst the reports are catalogues -- they make great reading for someone interested in buying an idea from those presented, but they don't make thrilling reading for the uninitiated. At the same time, they often assume knowledge of various data series that is not possessed by many outside experienced statistical system staff. As a corollary, some fail to cite relevant research literature outside that produced within the statistical system.

Part of these features may be a matter of choice of audience. I have assumed that the desired audience consists of both Federal Statistical System staff and researchers in related fields from academia and commercial domains. The government, academic, and commercial research sectors have much to gain from learning about each others methods. The paper series could be enhanced by seeking input from the two other sectors. At the very least, this might entail a forced literature review within each paper; at a higher intensity this might involve the subcommittee membership of those outside the Federal system. Even the input from outsiders may not sufficient.

Figure 3

Topics of Statistical Policy Working Papers

Multiple Error Sources	3 - CPS Error Profile 4 - Nonsampling Error Terms 13 - Federal Longitudinal Surveys 15 - Quality in Establishment Surveys
Coverage Error	17 - Coverage Error
Nonresponse Error	
Sampling Error	
Measurement Error: Interviewer	
Measurement Error: Questionnaire	10 - Developing Questionnaires
Measurement Error: Respondents	
Measurement Error: Mode of Data Collection	6 - Uses of Administrative Records 12 - Telephone Data Collection 19 - Computer Assisted Surveys
Processing	2 - Statistical Disclosure 5 - Statistical Matching 11 - Industry Coding Systems 18 - Data Editing
Estimation	7 - Time Series Revision

Topics Not Classifiable Easily in Error/Quality Terms

Topical focus	1 - Statistics for Allocation of Funds 16 - Reporting in Employer Data Systems
Administration	8 - Statistical Interagency Agreements 9 - Contracting for Surveys
Other	14 - Uses of Microcomputers

Missing Topics of Statistical Policy Working Papers

Coverage Error	Problems using households as sampling frame elements
Nonresponse Error	Combining social science and statistical models of participation
Sampling Error	Statistical software for estimation; generalized variance models; alternative estimators for public use files
Measurement Error: Interviewer	Training; variance models; reinterview programs; monitoring of telephone interviewers
Measurement Error: Questionnaire	Developmental methods in cognitive laboratories; pretesting regimens; imbedding experiments in surveys
Measurement Error: Mode of Data Collection	Mail and self-administered surveys; mixed mode surveys
Processing	Statistical quality control; automated coding
Estimation	Model-based Estimation

2.2. OMB Series as Cross-Fertilization Among Federal Statistical Agencies

In my fifteen years of working with Federal statistical agencies from my academic base, I was consistently reminded of the relative isolation of individual agencies from each other. As most people in this room know, it is not uncommon for very similar lines of research and development to be pursued without much coordination across agencies. The arguments for this are that different problems faced by the agencies demand different solutions. The arguments against are that functionally equivalent solutions are often created by two different agencies at twice the cost.

The working paper series has had, I believe, a beneficial unanticipated effect at reduction on interagency duplication. First, the subcommittees consist of members from several different agencies. Second, the tasks of the subcommittees often involve collecting information from many statistical agencies. The members thereby learn of work going on in agencies they normally don't visit. Third, recommendations of the papers often seek to apply standards across agencies, and the committees are forced to face the difficulty of system wide standards.

This is laudable and necessary. Is it sufficient? Clearly not. That is, working subcommittees of the Federal Committee on Statistical Methodology are temporary, normally have an agenda limited to the report, and do not generally follow up on logical conclusions of the report. Our dispersed statistical system, with all the benefits that specialization offers, misses opportunities to implement recommendations of these working papers.

2.3. OMB Series as a Prod to New Developments

Several of the papers treat topics where only one or two agencies are making major contributions and most others fall behind. For example, the Time Series Revision paper, the industry coding paper, the paper on computer assisted surveys, all fall into this category.

If I can temporarily put on the hat of an OMB staff member, this perspective seems to be the most central to the goals of the group. If reports like this can serve to improve the quality of work ongoing in several agencies, investments by one agency might quickly reap benefits in many agencies.

Some of the reports are poised for such effects, but the statistical system seems to miss more opportunities than necessary. Interagency agreements can be forged to promote such technology transfer. That is, consultation or subcontracting can be obtained within existing regulations. However, this requires the target agency to acknowledge the need for such upgrading. Could OMB

facilitate this process? I am too naive to know, but the existence of a pool of funds at the OMB staff level to assure the spread of innovation across agencies through detail of staff and other mechanisms would be productive.

Are there areas of innovation that can profit from coordination? Certainly. The use of CATI/CAPI is one that comes to mind quickly. It is now an area in which separate expenditures are being made by several agencies, where no standards have been well-defined, where different solutions, with essentially that same cost/benefit structure, may evolve across different agencies.

The prod to new developments, however, demands that the papers end with a series of recommendations. The authors should stimulate the readers, dare I say, challenge the readers, toward improving current practice. After the detailed investigation needed for these reports, they are uniquely qualified to offer such recommendations. Only a minority of the reports end with such recommendations. This should be part of the charge to each committee.

3. The Need for a Structure of the Working Paper Series

As I age, I must admit that I find more appeal in structures that guide our research and development in survey design and implementation, as opposed to reacting to each new idea without an explicit framework. In the academic world major theories provide that structure; they help to identify what are the important questions; they guide the development of new ideas. The application of the word "theory" to social and economic data production is rare. We do work that is guided by statistical theories, social science theories, organizational theories, and computer science theories. We are, however, basically on the applied side of research and development. We have a data collection and estimation vehicle (e.g., a survey) which is used for many substantive purposes. We are interested in knowledge that improves the vehicle and less interested in anything else.

As I understand the Federal Committee on Statistical Methodology, the topics for papers are essentially the fruit of discussions of the committee members. This is fine for assuring interest in the paper series among subcommittee members, but fails to assure coverage of important topics. I have suggested a total survey error structure above. The reports should have both measurement and reduction of error in mind. The widely perceived worth of sampling error as a criterion of evaluation of data owes its existence largely to well accepted estimators of the error. We currently lack comparably well accepted measures for nonsampling errors, but the report series could be used as a vehicle to stimulate such measures.

Finally, another way to structure the report series is around major problems facing the Federal statistical system in the near and far term (see Figure 4). These, in my view, should form the core attention of the working paper series. The first I mention may be the most controversial. The statistical literature on survey design is schizophrenic on costs. On one hand, there exist models which demonstrate that only through knowing cost components can design optimization be achieved. On the other hand, there is little serious treatment of survey costs by statisticians or those from other disciplines.

Figure 4

Likely Problems Facing Federal Data in the Near/Far Term

1. Identification of cost components associated with error-related design features
2. Integration of question changes motivated by cognitive research into ongoing surveys
3. Public cooperation with data collection requests and coverage of subpopulations on sampling frames
4. Development of mixed strategy designs, tailored to diverse subpopulations
5. Development of nonsampling error indicators; implementation of statistical quality control procedures
6. Training of statisticians and social scientists in survey research; recruitment/retention of trained staff

The second issue has both a restrictive and more global meaning. First, the work ongoing in so-called cognitive laboratories is seeking to identify principles influential of measurement error in question-answer sequences. The Federal statistical system at the current time has no good mechanism for the orderly introduction of change in questionnaires. For the vast majority of ongoing surveys, questionnaires remain static despite evidence of improved alternative measures. The value of unbroken time series and the assumptions of canceling biases in over-time comparisons are used to justify inactivity. Americans have very interesting reactions when they visit Cuba or see scenes of the country. They marvel at the maintenance of U.S. manufactured cars in their original state from the 1950's. They are at once proud of the ongoing use of older vehicles and humored by the lack of progress. A U.S. auto manufacturer would quickly go out of business if he were continuing to market 1950's designs. Indeed,

the watchword in that industry in continued investment in change, designing systems to permit ongoing change, making change part of the design. Survey researchers are driving 1950's vehicles in the 1990's. What we dearly lack is the will to mount ongoing programs of ongoing improvement in data series.

The third likely issue of import is the role of voluntary participation in surveys over the coming years. Some countries in Western Europe have experienced political shocks to response rates (e.g., Sweden, West Germany). Public debate about surveys in these countries has led to lower cooperation with survey requests. In some cases documented effects on survey statistics exist. That is, the nonresponse error becomes visible to even the most naive reader of statistics. At this point, there was little the researchers were prepared to do in terms of reaction of field interviewers or construction of adjustment schemes. We must acknowledge that public cooperation is a fragile base on which the scaffolding of inference lies. To improve participation or to adjust inference in the presence of lower participation, understanding of the decision to participate must be obtained. This is an issue that faces the entire statistical system, indeed, the entire industry of information collection.

The fourth issue is not unrelated to the problems of participation. As the diversity of the U.S. population increases, survey designs that tailor procedures to different subpopulations grow. Large portions of the population remain covered by traditional frames, cooperative and competent to provide information using cheap data collection methods. Others fail to be covered on traditional frames, have difficulty providing information, and fear harmful consequences from their participation. The coming years are likely to find greater appeal in mixed design strategies -- multiple frames, multiple data collection modes, tailored questionnaires to subpopulations. The models exist in the survey design literature, but they need careful attention.

The final problem listed above concerns a crisis looming ahead for the social measurement industry in this country. Like all endeavors that require quantitative literacy social and economic statistics are currently facing a shortage of qualified personnel. If this were not bad enough, we also suffer from a worse problem -- the absence of ongoing training programs. It's not merely that students aren't entering the field; it's not clear how they can within traditional academic programs. Let's examine the problem. Sampling statistics was well developed by the early 1950's; it is not a "hot" area of development, attracting the best and brightest of students. Instead, a variety of analytic statistical developments are more emergent. Young Ph.D.'s labelling themselves as sampling statisticians are unlikely to have an easy route to tenure in an academic department. Within the social sciences the difficulties might be greater, with great pressure on students to

develop areas of expertise which are central to the dominant paradigms in the discipline. Survey methodology is not one of them in any discipline. There are two results of this: 1) a gross inadequacy of training of new staff coming into the statistical system in topics relevant to survey quality. (This is not a comment on their training as statisticians, psychologists, or economists.) and 2) a reduction in the number of academic researchers devoted to the craft of social measurement. There is a clear conclusion here: the statistical system has to get serious about training of staff it needs for the future. This means support of specialized graduate programs, focused continuing education, onsite training, and other similar mechanisms.

The two types of structure - quality/cost components of data series and problems facing the system - suggest two paper series, one devoted to technical issues, another to administrative and professional issues.

4. Other Comments, Not Elsewhere Classified

I must admit confusion about the term, "working paper series." In an academic setting this term is used to describe papers in the process of being refined or papers not worthy of being refined. People are sometimes "working" on them. The better ones change over time, they evolve to a better state. This doesn't seem to fit well with the OMB Working Paper Series. Most all remain in their original state.

I don't want to change the name of the series; I'd rather see the series periodically updated. Several of the papers were valuable only for a short period of time (e.g., microcomputers; telephone data collection). Having a well-defined structure to the series might define a set of ongoing updates of papers devoted to individual topics.

There is another connotation of "working" when attached to paper series. That is, they are "working" toward quality improvements in the statistical system. I like this connotation. But it implies two burdens not uniformly accepted: a) a set of recommendations at the end of reports, b) follow through by OMB or individual agencies to implement change. On this definition, I think, the paper series has not achieved full success.

Another problem with the series are the costs and benefits assigned to authors of the reports. Contrary to my colleagues in academia, statistical system staff rarely experience career-enhancing effects of writing such papers. There is the value of education about other agencies, of "networking" with other members of the statistical system, and of learning more about important issues facing the system. On the other hand, I've learned that this is work essentially performed at nights and weekends by people

already very busy. Now, night and weekend work is commonly very productive and I have no problem with such a plan. What I do regret (and think it bad for the health of the system) is that such work is given so little value by many of the home agencies. OMB might consider remedying this with some more formal recognition of the writers of these reports. At the very least, the authors of the report might be given a more prominent position on the covers of the papers.

It strikes me that this seminar is an ideal forum for generating discussion on the future of this series. I recommend several questions:

Have the basic issues changed since the report?

- because of the paper?
- in spite of the paper?

Is it time to redo the paper, to update it?

Are there subtopics now of sufficient importance that they deserve separate treatment?

5. Personal note

This working paper series consistently contains the name of one person, from the first to the last - Maria Gonzalez. The Federal Statistical System often focuses its attention on data series structures and organizations, not people, but the success of any endeavor that spans decades depends on key people. In this paper series the key person is unambiguously Maria. As those of you who know her well can attest, she has been a rock of rationality, courtesy, integrity, and absolute honesty in her work on the Federal Committee on Statistical Methodology. She alone can succeed in pressing overworked federal statisticians to take on projects for the benefit of the whole system. Her near unique ability to suggest ideas in a manner that allows the hearers to believe they are their own ideas is a marvel. Her perseverance toward important goals of quality improvement and coordination have made the working paper series and this conference possible.

Session 1
SURVEY QUALITY PROFILES

THE SIPP QUALITY PROFILE

Thomas B. Jabine
Statistical Consultant

A. Introduction

The Survey of Income and Program Participation (SIPP) is a longitudinal national household survey which has been conducted by the U.S. Bureau of the Census since 1983, following several years of developmental research. The goal of the survey, which uses a rotating panel design, is to provide policy makers with comprehensive and accurate data about the levels and determinants of the income of U.S. persons and households and about their participation in a broad range of income transfer and welfare programs.

The SIPP quality profile summarizes current knowledge about the sources and magnitude of errors based on SIPP. An initial version of a SIPP quality profile was issued in 1987 (U.S. Bureau of the Census, 1987) and an updated and expanded version was prepared in 1989 (U.S. Bureau of the Census, 1990).¹

This paper describes the purposes of developing a quality profile for a survey or other statistical program and the process of preparing and updating a quality profile, using the SIPP Quality Profile as an illustration. The contents of the updated version will be discussed briefly. Those who wish to evaluate the quality of SIPP data on specific topics or to develop an overall judgement about the quality of SIPP data are referred to the latest version of the SIPP Quality Profile and the other sources of information that it identifies.

Section B outlines the development of the quality profile concept and identifies some publications of the last 4 decades that could be regarded as forerunners of the current model. Section C explains the origin of the SIPP Quality Profile. Section D provides an overview of the updated version: its intended audiences, purposes, sources of information and structure. The contents are discussed briefly in Section E. In the concluding section, I discuss the role of a quality profile in the broad context of survey quality control and improvement.

¹ For a copy of the latest version, write to Dr. Daniel Kasprzyk, Chief, SIPP Research and Coordination Staff, Office of the Director, Bureau of the Census, Washington DC 20233.

B. Some Forerunners of the Quality Profile

The theoretical foundation for a quality profile rests on various models that have been developed for the measurement and analysis of errors in surveys, especially the Census Bureau model, which integrates components of sampling and nonsampling error and the interactions between them (Hansen, Hurwitz and Berstad, 1959). Dalenius (1974) formalized the concept of total survey design, using the Census Bureau model to guide the allocation of resources to minimize total error in a survey.

Based on this foundation, there have been several broad qualitative and quantitative reviews of the quality of data from censuses and surveys, featuring direct and indirect data about the various components of error. Zarkovich (1966) published what was perhaps the first systematic treatment of nonsampling errors in surveys, with emphasis on procedures for their measurement and control, and including numerous examples of specific information about nonsampling errors from surveys and censuses in many countries. Bailar and Lanphier (1978), in a pilot test of methodology for the evaluation of survey practices, reviewed the quality-related design features of 36 U.S. surveys. Their review was not based on direct measures of errors, but the frequency with which they found indirect evidence of low quality was high enough to be disturbing and to suggest a need for greater attention to the quality of survey designs and practices.

A United Nations (1982) manual on Nonsampling Errors in Household Surveys, prepared for use in developing countries, systematically explores the different sources and types of nonsampling error and provides illustrative data from numerous household surveys throughout the world. Statistical Policy Working Paper 15 (Office of Management and Budget, 1988) performs a similar function for Federally sponsored establishment surveys in this country.

Compilations of information about the quality of surveys have two main audiences: survey designers/managers and users of survey data. To ensure that the latter have access to such information, standards have been developed for the dissemination, in survey publications, of information about errors. An early example of such standards was Census Bureau Technical Paper 32 (1974). Today, several Federal statistical agencies apply similar standards in their publication programs.

There have been some publications devoted entirely to the quality of data on a specific topic in a census or survey. An early example was a detailed appraisal of the income data from the 1950 Census of Population (Conference on Research in Income and Wealth, 1958). The most immediate forerunner of the SIPP Quality Profile was Statistical Policy Working Paper 3 (Brooks and Bailar, 1978), which provided an error profile for estimates of

unemployment from the Current Population Survey (CPS). Jabine (1987) provided a detailed analysis of the quality of data on chronic conditions reported in the National Health Interview Survey.

There are two fairly evident differences between the CPS error profile and the SIPP quality profile. The most obvious is the switch from "error" to "quality" as the defining adjective for the profile's content. While this may seem to be only a semantic change, it reflects a feeling, undoubtedly shared by the authors of the CPS error profile, that the goals of such a publication are constructive. The use of the term quality seems more in keeping with today's emphasis on quality control and improvement in all kinds of endeavors, including surveys. The other basic difference is that the SIPP quality profile covers the quality of estimates for all of the topics included in SIPP, whereas the CPS error profile covered only one of the many topics included in that survey.

Other U.S. statistical agencies are undertaking similar although not identical efforts. The Energy Information Administration, for example, periodically publishes reports in a series called An Assessment of the Quality of Selected EIA Data Series. These reports rely largely on the technique of comparing data from EIA surveys with more or less comparable data from other sources and analyzing the differences that are observed. Janet Norwood, in a paper presented at the Census Bureau's Third Annual Research Conference, stated that the Bureau of Labor Statistics was planning to develop a comprehensive error profile for each of its surveys (Norwood, 1987, pp. 217-218).

C. Origin of the SIPP Quality Profile

The SIPP is a major longitudinal survey. The start of the survey was preceded by several years of research and development, an effort known as the Income Survey Development Program. The evolution of SIPP's complex survey design did not end when the survey became operational late in 1983. Methodological research and evaluation studies have continued at a substantial pace and the results of these studies, along with accumulated performance statistics, feedback from users and adjustments made necessary by reductions in funding, have led to significant changes in the survey design and procedures. Thus, SIPP is still in the early stages of its evolution, in contrast to the Current Population Survey which, although not immune to evaluation and improvement, has reached a more mature and stable phase.

In 1984 the Social Science Research Council and the Survey Research Methods Section of the American Statistical Association, with the encouragement and support of the Census Bureau, established a Working Group on the Technical Aspects of SIPP to provide

advice to the Census Bureau on research priorities and the translation of research findings into changes in the survey design and procedures. (The Social Science Research Council later relinquished its sponsorship role.) An early recommendation of the Working Group was that the Census Bureau prepare a compendium of research results and other information about the quality of SIPP data. Members of the Working Group believed that a systematic account of information about the different kinds of errors that affect estimates from SIPP would be invaluable as a guide in setting research priorities and applying the principles of total survey design to SIPP. Given the substantial amount of ongoing research, they recommended that such a quality profile be updated periodically, perhaps every two years.

The Census Bureau accepted the Working Group recommendation and produced the Quality Profile for the Survey of Income and Participation (King, Petroni and Singh, 1987), early drafts of which were reviewed by several members of the Working Group. New information continued to flow in at a rapid rate and toward the end of 1988, Census decided that it was time to start work on an update. The updated version, published in mid-1990, was prepared by the author of this paper with substantial assistance from Karen King and Rita Petroni of the Census Bureau's Statistical Methods Division. Although the general structure of the two versions is similar, the update contains much new material and some of the earlier sections were significantly revised. It also includes an index. The new version benefitted from reviews by several members of the SIPP Working Group and Census staff. Special thanks are due to Daniel Kasprzyk and Rajendra Singh for their support of the project.

D. Overview of Version 2

The SIPP Quality Profile is intended to serve two main audiences: "users of SIPP data and those who are responsible for or have an interest in the SIPP design and methodology." The interests of these two groups are different. Users want to know how the errors associated with specific categories or classes of data are likely to affect their analyses. SIPP designers and managers need to know the magnitude of errors associated with specific design features, in order to control the quality of the survey estimates and to guide the allocation of resources available for their improvement. Besides these two primary audiences, it was expected that the publication would be of interest to persons concerned with the design of longitudinal surveys other than SIPP and to two special groups: the ASA/SRM Working Group and a Panel to Evaluate the Survey of Income and Participation, convened by the Committee on National Statistics at the request of the Census Bureau.

Information about the components of error that affect SIPP data comes from four sources:

- o Performance statistics, such as unit and item non-response rates and reports based on quality control procedures used in data collection and processing operations.
- o Methodological experiments. Both in the developmental period and since the start of survey operations, there have been numerous methodological experiments involving design features such as length of questionnaire, respondent rules, use of respondent incentives, increased use of telephone interviewing and methods of adjustment for nonresponse.
- o Micro-evaluation studies. The outstanding example is the SIPP Record Check Study, in which individual survey responses to questions about program participation and benefits were compared with administrative data for each of several programs.
- o Macro-evaluation studies. There have been numerous comparisons of SIPP data with data on the same topics from other surveys, especially the Current Population Survey, and from program records.

Assembling the relevant documentation was a challenge. SIPP has probably generated more methodological documentation than any other survey that has been in existence for a similar length of time. The list of 161 references provided in the updated version of the Quality Profile, which includes only those items that were actually cited in the report, is nearly double the size of the list included in the first version. The most commonly used sources were: the SIPP Working Paper series; the annual proceedings of the Survey Research Methods, Social Statistics and Business and Economic Statistics sections of the American Statistical Association; the proceedings of the Census Bureau's Annual Research Conferences; and internal Census Bureau memoranda. The report informs readers how to obtain copies of any of the internal memoranda in which they are interested.

Finding a suitable framework in which to present all of this information about different components of error also presented a challenge. The traditional approach is to organize the material according to the main phases of the survey: sample selection, data collection, data processing and estimation. The core of the Quality Profile (Chapters 3 through 8) is, in fact, organized in that manner, with one chapter devoted to sample selection, three to data collection (covering data collection procedures, nonresponse error and measurement error) and one each to data processing and estimation.

Two important topics did not fit neatly within this framework. Chapter 9, Sampling Errors, covers the procedures used to estimate sampling errors and the relationship between sampling errors and sample size. Chapter 10, one of the longer chapters, is called "Evaluation of Estimates" and covers both comparisons of SIPP estimates with data from other sources and indicators of errors of undercoverage. The remaining chapters, 1, 2 and 11, provide an introduction, an overview of the survey and a summary, respectively.

The structure of the SIPP Quality Profile is similar to that of its chief forerunner, the CPS Error Profile. The main differences are the division of the material on data collection (called "Observational Design and Implementation" in the CPS Error Profile) into three chapters, and the addition of the chapters on sampling errors and evaluation of estimates.

Our goal was to provide, insofar as available, quantitative information about overall error and its components. Hence, the report includes 6 figures and 43 tables, a substantial increase over the number included in the first version. Space limitations preclude inclusion of tables in this paper, but for those who may be interested, the numbers of some key tables and figures from the publication are given in the following section.

E. Summary of Findings

Major sources of error

The SIPP Quality Profile does not contain any broad conclusions about how successful SIPP has been so far in fulfilling its goals. Our goal was to provide enough information about the quality of the survey data so that individuals and groups like the Committee on National Statistics Panel to Evaluate SIPP could reach their own conclusions. The summary chapter does, however, identify what stood out as the three main sources of error in SIPP estimates: nonresponse, differential undercoverage and measurement error.

As in any longitudinal survey, unit nonresponse increases in succeeding rounds (called "waves" in SIPP) of the survey. Table 5.1 (not included with this paper, see the report) shows the data available as of 1989 on unit nonresponse by wave for each panel of the survey (households and individuals in each panel are interviewed 8 or 9 times, at 4-month intervals). The rates are relatively low -- 4.9 to 7.6 percent -- for the first wave, but increase to over 20 percent at the final wave of each panel. This relatively high attrition is due in part to the difficulty of tracking households and individuals that move, as is required by the SIPP design. The characteristics associated with unit nonresponse have been analyzed in detail, and these analyses have

guided the development of estimation procedures designed to minimize the biases that result from differences between the characteristics of respondents and nonrespondents.

Item nonresponse has been low for core items on labor force activity, income reciprocity and asset ownership. It has been somewhat higher for income amounts, especially self-employment earnings and interest. In the topical modules (questions not asked in every wave), especially high nonresponse has occurred for questions on asset amounts.

Indicators of differential undercoverage in SIPP for population subgroups defined by age, race and sex are shown in Table 10.13 of the report. The table shows the reciprocals of the weights that are applied in order to make the simple unbiased estimate for each subgroup agree with an independent estimate that uses the Population Census count as a benchmark. The group most affected is young adult black males. The ratios for black females in the same age group are also quite low. At least for the males, the coverage ratios shown understate the amount of undercoverage, because the ratios do not include any adjustment for census undercoverage, which is known to be above average for this population subgroup.

Similar patterns of undercoverage have been observed in the Current Population Survey and other national household surveys. The second-stage ratio adjustments used for both cross-sectional and longitudinal estimates to compensate for undercoverage are believed to reduce both the sampling error and bias of the estimates. The effects of these adjustments on sampling errors can be estimated, but little is known about their effects on biases associated with undercoverage.

Measurement error takes many forms, but perhaps its most significant manifestation in SIPP has been the seam problem, i.e., a pronounced tendency for survey respondents to report month-to-month changes for months in adjacent waves at substantially higher rates than for adjacent months within a single wave. Figure 6.1 in the report provides a graphic illustration of the seam effect on reports of changes in earnings. Pronounced effects have been noted for most income reciprocity and amount variables. Because of the rotation group design used in SIPP, cross-sectional estimates of transitions are not likely to be seriously distorted by this pattern of reporting, but it can affect estimates of the covariance structure and may have adverse effects on multivariate analyses dealing with transitions or length of spells.

Table 6.6 in the report shows some early results from the SIPP Record Check Study. The sample sizes are small, and the table shows results for only two of the four states included in the study. For the State of Wisconsin, significant levels of underreporting were found for participation in two programs and

benefit amounts in one other program. The full results from the Record Check Study will provide the best direct information so far available on levels of measurement error in SIPP and will be a valuable resource for studying the sources and correlates of response bias and response error variance.

Current research

An active program of SIPP methodological and evaluation research is continuing. The main areas of research include:

- o The design of the questionnaires and the structure of the interviews. Laboratory research is being conducted to study the cognitive aspects of SIPP interviews and how they relate to seam effects and other kinds of reporting errors. Field experiments have been conducted to test the feasibility of providing feedback of prior wave information and encouraging greater use of records in interviews.
- o Interview mode. An experiment with increased use of telephone interviewing is being evaluated to determine whether to adopt the procedures that were tested. For the longer term, the Census Bureau is arranging for the development of a prototype questionnaire for use in computer-assisted personal interviewing (CAPI), in order to evaluate the potential effectiveness of this collection mode in SIPP.
- o Estimation procedures. The broad goal for this area of investigation is to develop estimation procedures for SIPP that make effective use of auxiliary data available from both the Current Population Survey and administrative records. An initial study of the feasibility of reducing variances by using IRS data as controls in the second-stage ratio estimation procedure showed considerable promise.

Research in these and other aspects of the survey is proceeding at a pace that suggests the desirability of preparing updates of the SIPP Quality Profile on a regular basis.

Areas of research that have been relatively untouched so far include the effects of interviewer variance and the conditioning effects of repeated interviews on response error. For the latter, the overlapping panel design used in SIPP offers the possibility of comparing cross-sectional estimates for households and persons that have been in the sample for varying lengths of time. There is also a need to update some of the earlier evaluation studies in order to monitor the effects of design changes since the beginning of the survey. Much of the research reported in versions 1 and 2 of the

SIPP Quality Profile, including the Record Check Study, which is the only source of direct information on the size of individual reporting errors, is based on data from the 1984 panel.

F. Conclusions

Judging from some comments by users of the initial version and reviewers of the preliminary draft of the updated version of the SIPP Quality Profile, the systematic compilation and publication of information about the nature and sources of error in a major continuing survey like SIPP, with periodic updates, is a worthwhile undertaking. A more definitive evaluation of its utility will be possible now that the updated version has been published and is being widely distributed. The author believes that the preparation of quality profiles could be valuable in connection with efforts to track and improve the quality of data from other major continuing national surveys, such as the Current Population Survey, the National Health Interview Survey, the National Crime Survey, the Annual Survey of Manufactures and the Monthly Retail Trade Survey. The technique is applicable to both household and establishment surveys.

Maintaining and improving the quality of survey data is a never-ending job for survey designers and managers, and there is room for a multiplicity of approaches. Some Federal agencies are making a strong commitment to the application, to survey operations, of Deming's philosophy and techniques for total quality management. That approach implies not just measurement of errors and identification of their sources, but modification of the survey process as needed to eliminate or reduce the effects of significant sources of error. The other paper presented at this session (Hanuschak, 1990) provides an example of this model of survey quality management, with active participation and commitment to quality improvement by key managers in the organization. The same commitment to the quality of data can be seen in the work of the sponsors and participants in this conference and they deserve our thanks for it.

REFERENCES

Bailar, B. and Lanphier, M. (1978), Development of Survey Methods to Assess Survey Practices, Washington DC: American Statistical Association.

Brooks, C. and Bailar, B. (1978), An Error Profile: Employment as Measured by the Current Population Survey, Statistical Policy Working Paper 3, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce.

Conference on Research in Income and Wealth (1958), An Appraisal of the 1950 Census Income Data, Studies in Income and Wealth, Vol.23, National Bureau of Economic Research, Princeton: Princeton University Press.

Dalenius, T. (1974), Ends and Means of Total Survey Design, Stockholm: University of Stockholm.

Energy Information Administration (1983), An Assessment of the Quality of Principal Data Series of the Energy Information Administration (first in a series of "state of the data" reports), Publication DOE/EIA-0292(82).

Hansen, M., Hurwitz, W. and Bershad, M. (1959), "Measurement Errors in Censuses and Surveys", Bulletin of the International Statistical Institute, 38:359-374.

Jabine, T. (1987), Reporting Chronic Conditions in the National Health Interview Survey: A Review of Findings From Evaluation Studies and Methodological Tests, Data From the National Health Survey, Series 2, No. 105, National Center for Health Statistics.

Jabine, T., assisted by King, K. and Petroni, R. (1990), Survey of Income and Program Participation: SIPP Quality Profile, Bureau of the Census, U.S. Department of Commerce.

King, K., Petroni, R. and Singh, R. (1987), Quality Profile for the Survey of Income and Program Participation, SIPP Working Paper No. 8708, Bureau of the Census, U.S. Department of Commerce.

Norwood, J. (1987), "What is Quality?" in Proceedings, Third Annual Research Conference, Bureau of the Census, U.S. Department of Commerce: 215-222.

Subcommittee on Measurement of Quality in Establishment Surveys (1988), Quality in Establishment Surveys, Statistical Policy Working Paper 15, Statistical Policy Office, U.S. Office of Management and Budget.

United Nations (1982), Non-sampling Errors in Household Surveys: Sources, Assessment and Control, UN Publication DP/UN/UBT-81-041/2, National Household Survey Capability Programme.

U.S. Census Bureau (1974), Standards for Discussion and Presentation of Errors in Data, Technical Paper 32, U.S. Department of Commerce.

Zarkovich, S. (1966), Quality of Statistical Data, Rome: Food and Agriculture Organization of the United Nations.

INITIAL REPORT ON THE QUALITY OF AGRICULTURAL SURVEY PROGRAM

George A. Hanuschak
National Agricultural Statistics Service

I. Background and Introduction

In December 1988, the National Agricultural Statistics Service (NASS) formed a Survey Quality Team (SQT) for its Agricultural Survey Program (ASP). The ASP is a series of integrated multiple sampling frame (area and list) based surveys throughout the agricultural calendar year. Some major items on the surveys are planted and harvested crop acreages, hog, cattle and sheep inventories, crop yields and production and on-farm grain storage. There was a major survey redesign from individual MF surveys to an integrated multiple frame survey program which was implemented over several years (1984 - 1986). The mission of the Survey Quality Team is to identify and develop statistical process control (SPC) methods for the management of the integrated Agricultural Survey Program. The SPC methods are based upon the fundamentals of total quality management (TQM) techniques developed by Edward Deming, Joseph Juran, Philip Crosby and other well-known TQM developers in the TQM and SPC literature. However, since much of the literature refers to "manufacturing" situations, it was adapted to fit the government agricultural survey situation. Several papers by Ron Fecso developed the basic model of survey quality used by the SQT. The first major milestone of the SQT was to be the development of a baseline "state of the survey" quality report.

The mission of the SQT is quite broad, challenging and critically important to the Agency's long term goal of routinely and continually improving survey quality. The team and the Agency also face this challenge in the light of severe budget pressure, in general, on Federal Statistics programs. However, the team feels that TQM and SPC methods are quite powerful tools, when properly applied, that can aid in measuring and improving survey quality over time.

One of the first lessons of total process control is to define the major steps in the total process. In the case of the ASP, one needs to first define or identify the major steps or stages of the ASP surveys. The survey quality team had identified the following steps (Exhibit I) as the major 22 processes of the survey. Unfortunately, each one of these survey stages or processes is probably susceptible to some type of errors or biases. The SQT developed the following profile (Exhibit II) of 24 potential sources of error or bias in the ASP.

Like any good statistical organization, the Agency has tried to minimize the probability of various nonsampling errors occurring

in the survey process. Controls include training, survey manuals and instructions, Agency Policy and Standards Memorandum, quality control checks on enumeration, reinterview studies, etc. Controlling and measuring nonsampling errors for a complex survey process will remain extremely challenging even with the best efforts at statistical process control. However, in the remainder of this report, the SQT defines and demonstrates how to use statistical process control and total quality management techniques to reduce total survey error over time.

Exhibit I - Major Survey Stages

Survey Clearance

Area Sampling Frame

(Construction, Maintenance and Sampling)

List Sampling Frame

(Construction, Maintenance and Sampling)

Survey Specifications

Design of Questionnaires

(Design, Print and Distribution)

Preparation of Manuals

(Interviewers, Supervisory and Editing)

Prepare Survey Software

(Data Entry, Survey Coordinator, Edit, Analysis, Summary, Data Base, Mail and Maintenance System, Etc.)

National/Regional Training Schools

Survey Management - Headquarters and State Statistical Offices

(Coordination of Procedures)

Presurvey Coding/Handling/Processing by State Statistical Offices

State Training Schools

Data Collection

Data Collection Quality Control

Manual Data Review and Coding

Data Entry and Validation

Data Edit and Review

Imputation, Analysis and Summarization

State Statistical Office Review of Survey Results

(including submission of estimates)

Headquarters Review and Release Preparation

Post Survey Updating

(Data Base and List Sampling Frame)

Post Survey Evaluations

Survey Research

Exhibit II - Some Potential Sources of Total Survey Error
in the Agricultural Survey Program

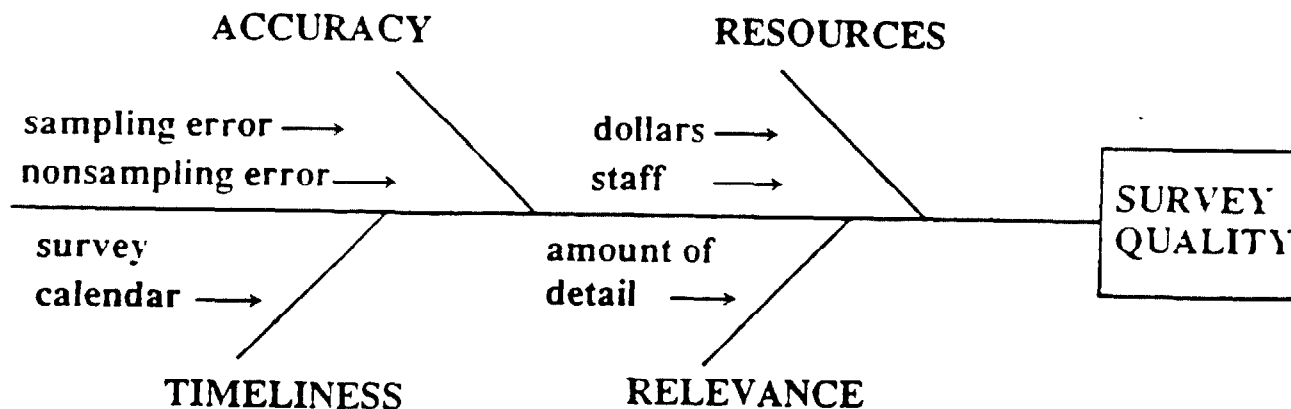
Undetected List Sampling Frame Duplication
List Sampling Frame (Old or Incorrect Control Data)
List - Undetected Reporting Duplication or other
reporting/enumeration errors or bias
List Sources of Questionable Quality used for List Sampling Frame
Build/Maintenance
Area Sampling Frame (Outdated Land Use Stratification)
List Sampling Frame (Any large operations not covered by the
frame)
Area Sampling Frame (Outdated Sample Segment - Aerial
Photography)
Different Farm Operation Description Questions
on Different Questionnaire Versions
Incorrect overlap/nonoverlap Determination
Incorrect Exception Report Handling (One Type of Survey Weighting
Factor)
Incorrect Coding (List Adjustment Survey Weighting Factors,
Completion/Imputation Codes, etc.)
Undetected Data Entry errors (pass all the way through the
editing system)
Shift in Mix of Data Collection Modes (Telephone, Computer
Assisted Telephone, Mail and Personal)
Shift in Mix of Respondents (Operator vs. Spouse vs. Other)
Incorrect Survey Master Records
Questionnaire Design (or Print) Errors
Unmeasured Major Changes in Survey or Estimation Procedures
(Headquarters or State Statistical Offices)
Error in Known Zero Determination (Is Respondent Validly Out of
Business?)
Overediting/Underediting of Survey Data
Potential Bias in Manual or Machine "Imputation" Procedures
Lack of Formal Outlier Handling Procedures (Non Robust or Non
Smooth Time Series Estimation)
Survey Processing Software
Shifts in Characteristics or Skill Level of Work Force
{(Enumerators, Statisticians, Programmers, Support Staff)
Experience in their current job, survey procedures
knowledge, farm knowledge, statistics knowledge, technology
skills, etc.}
Farmer or Respondent's level of understanding or grasping of
survey reporting concepts and item definitions (Cognitive
aspects).

II. The Components of Survey Quality

When faced with the problem of measuring and improving the quality of the ASP, one should consider the components of survey quality. Listing the components defines exactly what is meant by the term "survey quality" and highlights specific sub-areas that need to be explored.

Figure 1 shows the components of survey quality. It was developed by the Nonsampling Errors Research Section in the Survey Research Branch of NASS and adopted by the SQT. There are four major components related to survey quality -- accuracy, resources, timeliness, and relevance.

Figure 1 The components of survey quality



Accuracy is the component that first comes to mind when thinking about survey quality. NASS wants the survey indications to be as accurate as possible. Not only should the sampling errors be small, but also the nonsampling errors should be minimized. In large-scale surveys the relative sampling errors can be smaller than the relative size of the nonsampling errors. Factors such as undetected list sampling frame duplication, nonresponse, questionnaire wording, mode of interview, change in respondent, etc., can lead to substantial nonsampling errors.

The second component of survey quality is resources. Even if a survey organization can control the sampling and nonsampling errors, its ability to do so will be affected by the amount of dollars that are available to spend on the survey. The amount of dollars has a direct impact on sample sizes, list frame quality, pretesting, reinterview projects, editing programs, summary programs, analysis, etc. Also important is the amount and quality

of staff hours that can be devoted to a survey. Staff hours are affected by salaries, training, hiring practices, long-term career development, and organizational climate; components that are also greatly affected by the amount of dollars available. Most people quickly realize that the crucial problem is to take the fixed set of available resources and use those resources in a way that maximizes the survey quality.

The third component is timeliness. Of course, time could be considered another element of resources -- like dollars and staff. However, timeliness needs to be considered a component by itself because timeliness is crucial in the survey process. The impact and usefulness of survey indications are greatly affected by whether the survey data were collected one month or one year earlier. NASS has always stressed the need to collect data quickly and to release estimates as close to the survey reference date as possible. Thus, the survey calendar -- which is used to time all the steps of the survey -- is important to the survey quality.

The final component is relevance. Relevance is dependent on the needs of the users of NASS statistics, and those needs change from day to day. It is useless for NASS to collect a high-quality piece of information on farming if that piece of information has no relevance for the users of NASS statistics -- that piece of information simply becomes a product without a buyer. NASS must constantly assess the needs of people using its statistics to make sure that the collected information is relevant. The second aspect of relevance is internal to NASS. An example of internal relevance is whether the Agency wants direct expansion (level) or ratio (percent change) or both types of estimators out of the ASP.

III. Accuracy of Survey Soybean Acreage Estimates

NASS has an expert panel of Agency statisticians called the Agricultural Statistics Board (ASB) which reviews all survey indications (often multiple indications for any one item), and administrative or check data (such as the amount of soybeans crushed in processing plants) and adopts or sets the official estimates to be published.

Two concepts need to be defined - use and fitness. The ASB's use of the ASP indications was chosen as the primary "use" of the ASP. "Fitness" for use is evaluated by setting a standard for use and measuring adherence to the standard.

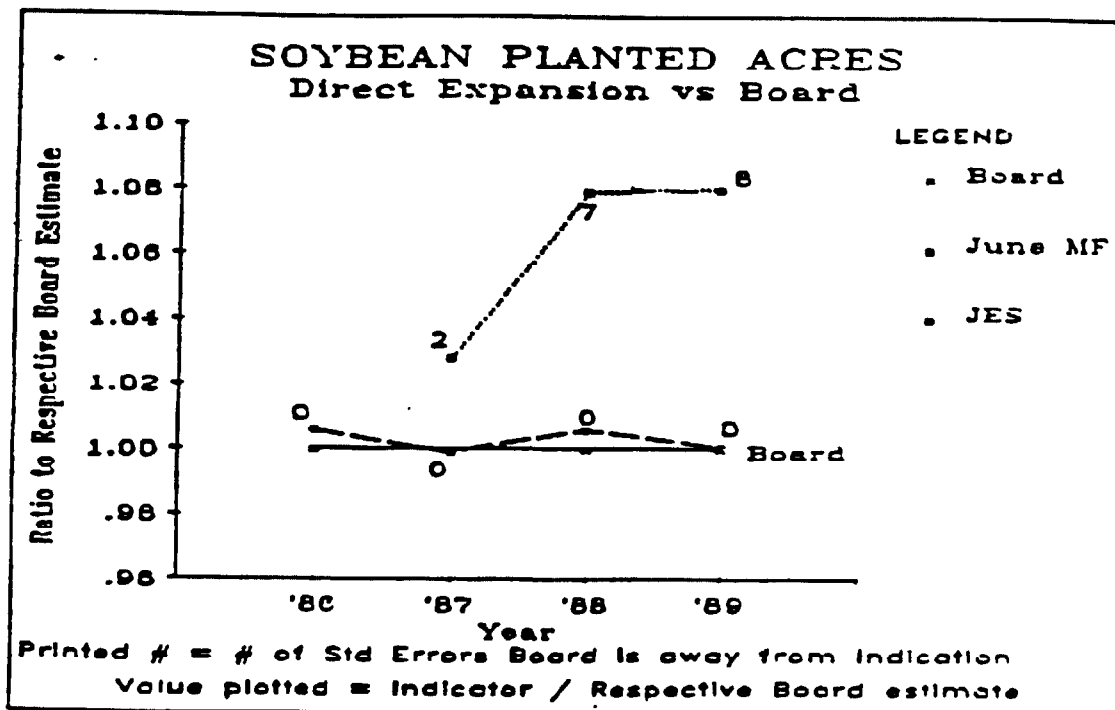
Ideally we would have standards for all the components of mean squared error (MSE) for the various commodity indications and administrative data used by the ASB. This would provide the ability to create statistically well defined composites of the data for use as the Board estimate or forecast. As this time we have measures of the variance for most indications, but have only enough

information about MSE's to recognize the importance of developing more extensive MSE measures. This section will provide information for Agency management to assess which areas are most in need of further study or research and/or corrective action.

The ASB's specific need is to have indications which serve as a solid basis for the official numbers. The following chart on soybean planted acreage display the degree to which the ASB has found the ASP indications to be "fit for use."

In reviewing the soybean planted acreage chart on ASB use you will observe the following:

1. The Agricultural Statistics Board finds the area sampling frame based June acreage estimate quite "fit for use."
2. The ASB does not find the integrated multiple frame based June acreage estimate "fit for use." It has an observed substantial upward bias which also changed substantially in magnitude between 1987 and 1988 and stayed at the larger magnitude in 1989 and 1990. Using Pareto analysis and an expert panel using TQM principles applied to surveys, the SQT identified the major suspected causes of the upward bias in the multiple frame based soybean acreage estimate. These suspected causes are:



1. Different Data Collection Methodologies

The area frame based acreage estimate is based upon a sample of about 16,000 sample segments throughout the U.S. Data collection is done completely by personal interviews using an aerial photograph to locate each crop field and recorded on a questionnaire by the interviewer with the farmers direct participation. Crop acreage data is collected and edited field by field. Farmers are probed to report waste acreage for each field. There are also five specific questions related to defining land operated now to which all the rest of the questions relate to.

On the integrated multiple frame survey, the majority of data collection is done by telephone (both conventional and computer assisted). The crop acreage data is collected for the entire farm (not field by field). Therefore farmers are probed for waste acreage only once, at best, when reporting crop acreage. There is no photographic aid for the farmer to refer to. There is only one or two questions on defining land operated now.

2. Undetected List Sampling Frame Duplication

There are sophisticated record linkage tools to identify and remove duplication on the list sampling frame. However, due to clerical resource constraints and funding to call farmers to resolve differences and the use of multiple list sources some duplication remains. A special study was designed in 1989 to measure remaining duplication and the effect on the estimates. The study showed that approximately 10 percent of the acreage difference was due to obvious list frame duplication.

3. No Formal Documented Outlier Handling Procedures

While there are several good analysis tools to identify outliers, there is no formal procedure for handling them. The area frame based acreage estimator is quite robust since the average expansion factor is about 200 and the segment size is 640 acres putting an upper bound on "influential observations". For the list sample, expansion factors are considerably larger and farm size does not have much of an upper bound. Thus it is much easier to get highly influential observations in the list sample. Development of a formal robust estimator for the list sample is highly recommended.

4. Different Imputation Methodologies

There are also different imputation methodologies. All imputation for the area frame is done manually by interviewers observations or statisticians. In the case of crop acreage if a farmer refuses the interviewer can still observe most of the crop fields and the crop. On the list sample, the imputation is a computerized algorithm that uses other reported survey data and list frame control data to impute for nonreported data cells.

5. Undetected Reporting Errors

Since the questionnaire design is different the undetected reporting error structure may also be different. For example, the screening questions on land operated on the area side are more detailed than the list questionnaire and may do a more accurate job of screening out landlords who are not active farmers at survey time. New farm programs may have also led to the formation of more complex farming operations, which may involve a different reporting error structure also.

6. Different Ratio Type Information and Sample Designs

On the area frame sample there is an 80 percent overlap from one year to the next. On the list frame sample (independent from year to year) there is negligible overlap. Thus the area frame sample also provides a paired sample ratio estimator.

It is important to note that there have also been two rather independent sources of data available to the ASB which also support following the area frame level. These are a Landsat satellite based regression estimator (1980-1987) which for major soybean states had variances at least twice as small as the direct expansion estimator but also were unbiased when compared to the ASB and direct expansion. The second source is the calculation of a soybean balance sheet which the ASB uses as an evaluation tool. A balance sheet takes the carryover from one crop year to the next and adds crop production to that and then subtracts crop utilization including exports from it to get a current balance. These balance sheets also support the area frame based crop acreage level. Thus the agency has attempted to verify the correct crop acreage level using several methods and independent data sources.

Even though there is an observed upward bias in the integrated multiple frame estimator for soybean acreage there are reasons for keeping it and reducing the bias. These reasons are:

1. Later crop season yield and production estimates are tied to the integrated multiple frame (IMF) approach.
2. State and sub-state level estimates from the IMF have much better precision than the corresponding area frame estimates.
3. Solving the bias problem associated with soybean acreage may well improve the entire IMF which is a survey 6 times a year with an average of 20-40 items (multivariate in nature). The Survey Quality Team has performed similar analysis for on-farm grain storage, and cattle and hog inventories. Some of the bias issues are item specific but others are associated with the total survey process or components of the survey process.
4. The IMF approach is substantially more cost efficient and involves less respondent burden than the area frame approach.

Most important is that the Agency is taking actions on all of these expected causes in 1989 and 1990. As previously mentioned there is now an improved list frame duplication adjustment procedure in place starting in June 1989. There is a reinterview research study being conducted in June 1990 to provide initial measures of previously undetected reporting errors. This study will involve the reinterviewing of a subsample of the list sample of farmers and record the crop data field by field and ask the more detailed land operated questions and compare the results. There are also research efforts underway to examine the imputation methodologies and to look at an across year design for list frame based estimators and evaluate several robust estimators. In addition the SQT has provided several quality measures to be monitored on the resource, relevance, timeliness and accuracy dimensions which should become operational in 1990-91.

The Agency is also developing alternative "proxies" to the true item values in addition to relying on the ASB process. An operational reinterview/reconciliation survey is being conducted in six major grain producing states in December 1990. There has also been an extensive operational soybean yield validation survey (198? - current) where farmers are asked to harvest specific fields and take just that grain to a grain elevator to be weighed and measured.

This "proxies" to true values are important in a survey evaluation program but are also complex and expensive to develop and implement.

As previously mentioned, use of earth resource satellite data has also been used by the Agency to develop more precise and accurate crop acreage estimates.

IV. Summary

It is the claim of the SQT that more consistent and timely process improvements can take place by using the principles of statistical process control and Total Quality Management. More formal survey quality measurement and monitoring mechanisms will provide the Agency's management with more and critically important information to manage the quality of the ASP. Also, most of these techniques will readily transfer to other survey programs in the Agency such as Prices Paid and Received by Farmers, the Farm Costs and Returns Survey, Objective Yield Surveys, Farm Labor Surveys, and even to new programs such as Water Quality and Food Safety Surveys, the National Animal Health Monitoring System and the Monthly Yield Survey Program.

There are several tools available for such a survey quality management system. First there are numerous charting techniques such as bar and pie charts for resource information, Board standardized indication graphs with standard errors, Gantt charts to display project management and survey schedule information, upper limit and lower limit control charts, multivariate control charts, Ishikawa fishbone diagrams and Pareto charts and analysis. Many of these were used in an earlier effort by the Nonsampling Errors Research Section when a statistical process control study was conducted on the Soybean Objective Yield Program.

Pareto analysis is one of the most powerful tools in quality monitoring systems. Pareto analysis ranks the potential errors in a system from most serious to least serious. The reasoning is that in many systems and not just surveys, there are a "vital few" and "trivial many" potential errors in the system. Thus, the most important beginning of evaluating the quality of a system is to identify where it is most likely to break down or fail. Once the ranking of potential errors is accomplished, then it is recommended to identify the allocation of resources for each potential error to see if management is allocating resources in a fashion that will truly minimize total survey error. Many Pareto analyses have demonstrated that the resource allocation was not in proper alignment with the true error structure.

Thus, more information on the true total survey error structure and appropriate resource allocations, is being provided to survey managers and administrators to form a basis for future improvements in total survey quality.

Considerable progress has been made by the Agency in addressing quality issues in its integrated multiple frame Agricultural Survey Program. Many of the discoveries will translate to improved quality on several other major Agency survey programs as well.

References

Beller, N., "Error Profile for Multiple Frame Surveys," Statistical Reporting Service, Research Report, 1979, Washington, DC.

Bosecker, R., "Integrated Agricultural Surveys," National Agricultural Statistics Service, Research Report No. SSB-89-05, Washington, DC, June 1989.

Fecso, R., "Survey Quality," Presented at the 2nd Quality Assurance in Government Symposium, Washington, DC, May 1989.

Fecso, R., Pafford, B., Tremblay, T., Johnson, R., "Quality Profile for Soybean Objective Yield Survey," National Agricultural Statistics Service, Unpublished Case Study, Washington, DC, 1988.

DISCUSSION

Barbara A. Bailer
American Statistical Association

I. What is a Quality Profile?

The first quality profile was called an error profile and it concerned the CPS employment statistics. To be more positive, error profiles have now become quality profiles. The purpose is to prepare a systematic and comprehensive account of survey operations, listing the operations, the potential sources of error, and how the error influences the uses of the survey statistics.

Quality profiles are still rare events. When asked why there are not more, survey producers have three main themes:

- ° The staff resources that would go into producing a quality profile are too great and are in competition with other, more urgent needs.
- ° Producing a report that tells about the errors in surveys would lead to less credibility in the statistics produced.
- ° Admitting that there are errors is admitting that we haven't done our jobs well.

In fact, there are many benefits to producing quality profiles. Some of these are as follows:

- ° to minimize total error, not just sampling error, within given cost constraints
- ° to force a thorough documentation of the survey process.
- ° to guide a user on the effects of possible errors and their impact on specific uses
- ° to develop a sound quality control program
- ° to use in training programs for new staff in either operations or research; and
- ° to use as the foundation for a sound research and analysis program

The development of a quality profile parallels the survey process and would contain the following elements:

1. Objectives and specifications of the survey
2. Sampling design and implementation
3. Observational design and implementation
4. Data processing
5. Estimation
6. Analysis and publication

Given this as my basic understanding, let me comment on the quality profile for SIPP and the quality assessment of the Agricultural Survey Program (ASP).

The two reports have some differences and some similarities. The SIPP profile summarizes what is known about sources and magnitudes of errors of estimates and addresses accuracy. The ASP report is written from the point of view of total quality management and uses many of the ideas of Deming, Juran, and Crosby. This report considers resources, timeliness, and relevance as major components of quality, along with accuracy. The aims of the two groups seem to be quite different.

The two reports each identify the same groups as their targets -- the users of the survey data outside the agency and producers of the survey inside the agency.

Another similarity is that both look at major phases of the survey operation, something essential for a quality profile.

A difference in the two reports was that the SIPP report actually identified four main sources of information on nonsampling errors:

- Performance data
- methodological experiments
- micro-evaluation studies
- macro-evaluation studies.

The ASP report was more concerned with process and how quality would be assessed. In fact, the report stresses the need not to identify too many sources of error because tracking everything down might take too long. Actually, I think the total quality management movement urges groups to use brainstorming techniques to identify all possible problems and then Pareto analysis to decide where to concentrate one's efforts.

Another similarity is that both reports left out major steps in the survey process. The SIPP report briefly listed the objectives of the survey, but said nothing about the objectives being conflicting. Producing a survey to give both cross-sectional and longitudinal data has been a new experience for the Census Bureau. The two objectives do conflict, at least from the resource point of view. There were some references to different needs in imputation, but the resource needs have probably had more impact on the survey.

The ASP report did not even list objectives of the survey as a potential source of error. Neither report really addressed the effects of staff training or compared the kinds of training, length of training, etc. It is fairly well known that performance data does not correlate well with interviewer performance on accuracy. Training could make a difference, but almost nothing is known at the present time.

Let me move now to some separate comments on the two reports, starting with the ASP report. There was a large group of people who worked on this survey quality team. Many of them have done excellent work in survey methodology, so I think we can expect great things from this group. The mission of the group is to contribute to NASS's long term goal of routinely and continually improving survey quality.

The focus on quality at NASS has taken on the language of the quality and productivity movement. For example, they use a simple definition of quality, "fitness for use." This led them on a search to decide what that meant and what objective criteria would be. Finally, they decided that they would measure it by comparison with the Agricultural Statistics Board (ASB) estimate. If the ASB value is within plus or minus two standard errors of the survey indication, then the survey indication is fit for use. And, in fact, they have five ratings: ideal, acceptable, workable, minimal, and out-of-control.

I find it hard to see why the Agricultural Statistics Board estimate would be used as the standard. In some cases, there are long time series and other indicators that the ASB uses to make its estimate. However, for some surveys they have much less information. Perhaps NASS is pushing the ASB to use the survey indicators or explain why they haven't. Though the example given in the paper about the integrated multiple frame based June acreage estimate was interesting, there will not always be that kind of other data available to compare with.

There is nothing about a Board estimate that measures accuracy. In some ways, it is as if the SIPP people looked at one of their macro indicators and said that if SIPP didn't come within two standard deviations of that estimate, then SIPP was not fit for use. At least, with a macro indicator, one might be able to untangle why estimates differ; that may not be possible to do with the ASB.

Following Deming's principles, I think the careful documentation of every survey for which millions of dollars are spent and on which important decisions are based is important to the profound understanding of which Deming speaks. A quality profile tells you what you know and what you don't know but should.

It was interesting to see that NASS also addressed resources, timeliness, and relevance as major components of quality. However, it was not clear how criteria would be set or measurements taken. The Gantt chart on the QAS was helpful in identifying time periods and overlaps of one round of survey with the next but it did not help individuals who have many surveys to work on identify overlapping periods of high intensity. The sentence "Too frequent use of overtime to correct a process that is out of control usually has a devastating effect on overall performance." What does out of control mean? How does it affect overall performance? How do you know these things unless you keep careful records on hours worked on a survey, overtime, and have some measure of a downturn in performance?

NASS has several good ideas about looking at relevance, timeliness, and resources as well as accuracy. It is an ambitious undertaking. I have one word of caution in their drive to use total quality management techniques to help them. They focus on several tools available for a survey quality management system including charting methods. I agree that these are useful tools. But what has been most helpful in the manufacturing and service industries where TQM is used is bringing in a team that has hands-on knowledge of all the facets of the survey. The team would include data collectors from states, edit specifications people, estimation people, those who set objectives. The tools would be something the team would be taught to use to help them. They would all need to learn basic concepts of variability. Only when all these people participate, do you get the profound knowledge that you need to improve a system, not merely tamper with it. As you recall, tampering with a system does not take care of the major changes needed to remove high variability due to special causes.

Let me now move to the SIPP report. This is a good report that gets periodic updating. There are areas not covered in the report, probably because they did not seem as urgent as the areas covered. However, I do believe that we will need to see a section on objectives, meeting multiple objectives, defining concepts, translating concepts into questions, and so forth. At the other end of the survey, something needs to be said about analysis and publication.

Though the Census Bureau does not use the language of total quality management, I know that they have thought along those lines. Using some of the performance measure standards flies in the face of everything Deming preaches. I'm talking about standards for response rates:

Outstanding.....	97.5 - 100.0
Commendable.....	95.5 - 97.4
Fully successful.....	91.5 - 95.4
Marginal.....	88.0 - 91.4
Unsatisfactory.....	87.9 and less

Instead of setting arbitrary standards for response rates and production, the Bureau needs to get a deeper understanding of what is possible in each type of area in which it does surveys. For example, response rates can be charted with upper and lower control limits for PSU's in New York City. Probably the response rates there very seldom, if ever, meet the commendable level. However, they may be within normal variability for that area. Only with positive efforts at changing the system can the response rates be lowered. This is partly what Dr. Deming thunders about -- blaming the worker who may be doing the best he or she can when it is the system at fault. Again, this labelling of people's work does not make the interviewer proud, and it is really tampering with the system.

The report gave lots of interesting information on household, person, and item response rates. Some of the non-response rates on asset data are such that it seems questionable that the survey is the right vehicle for collecting the data.

There is also emphasis on the seam problem, but this is nothing new. As I recall, it also showed up in the crime survey. It seems that certain biases are endemic to longitudinal surveys. So far the Bureau has been content to catalog the measured effect. We really need some creative thinking and some money to get some experiments going to look at recall errors, the displacement of events in time, and the time in sample problems. Though dependent interviewing may yield more consistent results, they may be no more accurate. Before action is taken to fix a problem, there needs to be a deeper understanding of why the problem exists.

There was very little information available on the extent of editing, what it does, why changes are made, and what we call editing and what we call imputation. Beller made some very pertinent comments in his 1979 error profile for NASS surveys. "The amount of editing on some questions resulted in changing the level of cattle and calves by an amount two or three times greater than the error caused by sampling. This amount of editing is cause for alarm in that it clearly shows a breakdown in the survey process." In both the NASS surveys and SIPP, we need to get a better picture -- a profound understanding -- of what editing is doing to the data.

One last point on SIPP. The only direct estimates of sampling error were for the third quarter of 1983 using 1984 panel data collected in wave one. The survey at that time was based on the 1970 census. It certainly seems time to recompute variances. Besides having incorrect variances, it seems like gilding the lily when the analysts are making actual and implied comparisons that they multiply by 1.6 times the standard error. The interpretations and the comparisons could be quite far off.

All in all, I enjoyed reading these papers. I think the documentation of SIPP is more complete but I think NASS is farther along in trying to improve quality. They do not want to document only; their real goal is improvement. I believe that is ultimately the SIPP goal too, but no strategy has yet been set forward on how to move in that direction.

DISCUSSION

Nancy A. Mathiowetz
U. S. Bureau of the Census

The data collected by Federal statistical agencies are used to both shape federal policy and change the distribution of federal expenditures; given the magnitude of the impact of these data, the need for high quality goes without question. In developing the Quality Profiles, the agencies responsible for this work are to be commended for continuing to move the discussion of error beyond that of sampling error and into the realm of the measurement of nonsampling error. Although most agencies have for years provided discussion of sampling error with release of their data and research findings, we are just beginning to develop a standard of reporting which includes a discussion of all of the components of total survey error.

Sources of Nonsampling Error

The sources of nonsampling error are many and include:

- the design of the study (e.g. longitudinal vs. cross sectional; length of recall period;
- the questionnaire, both the contents and the structure;
- the interviewer;
- the respondent; and
- the post-survey processing, including coding and keying of data.

Rather than reiterate issues raised in the Quality Profiles, I would like to suggest some other topics of investigation within these sources of nonsampling error. My goal in doing so, is not to criticize the work presented here, but to provide some ideas on where these Quality Profiles could be expanded.

Design

With respect to design, we still know little about the effects of longitudinal designs on the level of error and the error variance structure of reports over time. There has been research to indicate that respondents suffer from "conditioning" effects, that is the changing of behavior or the reporting of behavior in later interviews resulting from earlier interviews. Some conditioning may improve reporting in that the respondent knows

prior to the interview what are the nature of the questions; conditioning may also result in a reduction in reporting since respondents are now knowledgeable about the sequencing within an interview. In one study, the best predictor of error in reports of functional status in the fourth round of interviewing is the length of time it took to conduct the previous interviews. The finding suggests that conditioning effects may be reduced by something as subtle as reducing the length of an earlier interview. We need further research to understand how conditioning impacts the analysis of change over time and the structure of errors over time.

Longitudinal designs may also be affected by changes in the respondent, the interviewer, or even the interpretation and meaning of critical concepts in the questions, if the panel has a long life. With the proliferation of more longitudinal data collection efforts within the Federal Government, more research into what questions are sensitive and which are resistant to conditioning effects as well as which items are most affected by between interview changes, is necessary.

Questionnaire

As noted in a lecture to the Society of Government Economists, Janet Norwood stated that

...the quality of a statistical indicator is sometimes elusive and often difficult to define. Effective measurement requires an underlying conceptual framework and careful identification of the phenomenon to be estimated....

In the past 25 years, we have made great strides in understanding how sensitive response distributions are to minor changes in question wording. The merging of literatures from cognitive psychology, social linguistics, and social psychology with survey methodology has presented use with new means for attempting to reduce the levels of error associated with the questionnaire. What is now needed in the Federal statistical system is a means for evaluating the various forms by which the "same" information is collected and analyzed among various agencies. For example, in recent years, the proportion of individuals lacking health insurance has been a critical issue. The most widely cited data on insurance coverage comes from the Current Population Survey which asks whether each person in a household was covered at any time during the preceding year. Persons covered by any source at any time during the year are counted as insured. In 1987, the estimate for uninsured from the March CPS was 17.6 percent. Notice that this question asks whether the person has been covered "at any time" during the previous year. In contrast, questions from the 1980 National Medical Expenditure and Utilization Survey (NMCUES) and the 1987 National Medical

Expenditure Survey (NMES), both designed as one-year panel surveys, indicate that point in time estimates of the uninsured (at the time the person was interviewed) are approximately 14 to 16 percent at any one cross-section, but that estimates for all year uninsured are approximately 9 percent.

There is some conjecture that the response to the CPS may reflect a respondent's status at the time of the interview rather than in reference to any time in the previous year, due to the similarity in the estimates from CPS and the cross-sectional estimates from NMCUES and NMES. From a policy perspective the difference is critical -- whether to provide health insurance for the chronically uninsured, approximately 21 million people, or whether to provide insurance for all individuals ever uninsured, which appears to be approximately 35 million people in a given year. Those attempting to address this issue would benefit from a consistent definition of uninsured as well as a set of questions which asks about a consistent time period.

Interviewer

The use of response rates, hours per completed interview, and item nonresponse rates, traditionally used as measures of interviewer quality, only begin to capture the errors that are potentially associated with the interviewer's task. While each of these measures provides us with information that we believe is related to quality, we need to employ more measures that could be used with respect to understanding error for individual questions. How well do interviewers understand the concepts underlying the questions they are asking? Do they have sufficient training and understanding to ask non-directive probes when necessary to obtain an adequate answer? The increased movement toward telephone interviewing provides use with a means to routinely randomize interviews across interviewers to obtain measures of interviewer variance. We spend millions of dollars in the training of interviewers and yet know little about the most effective means for training interviewers or determining their ability to conduct the interview as trained. The review of one or more interviews by a supervisor provides some information, but if we believe that training interviewers to read questions exactly as written is worth the cost, we should be routinely evaluating the association between the delivery of questions and the error associated with the responses.

Editing and Coding

As noted in the SIPP Quality Profile, much of the between wave difference in industry and occupation appears to be a spurious result of either data collection or data processing. A similar problem can be found in the coding of medical conditions and

surgical procedures based on household reported data. Not only coding, but also editing procedures, can contribute to the overall level of error in estimates. For example, Duncan and Mathiowetz (1985), using microlevel validation data, found that trimming estimates of change in income between two years, that is disbelieving levels of change beyond a certain level as reported by household respondents, a procedure often done in editing data from longitudinal surveys of income, resulted in biased estimates of change and bias in the coefficients predicting income levels and change. Retrospective reports of income were more likely to be correct for those individuals with a large proportional change than for those with little or no change. The finding suggests that editing procedures should be conservative and based on empirically derived principles.

Whereas we have learned to be sensitive to question wording with respect to understanding potential sources of bias, and in doing so, demand documentation concerning question wording and study design, few, if any, studies provide information on effects of editing and coding processes. If consumers of the data are to understand all aspects of total survey error, coding and editing decisions need to be researched and documented.

Adjusting for Nonresponse

For the most part, nonresponse adjustments are made using demographic and segment information and little if any information concerning the nature of the nonresponse is factored into the adjustment. There is a growing body of literature which suggests that using information from call records, specifically separating refusals from those you were unable to locate, in a nonresponse adjustment may prove beneficial, since difficult to locate (but eventually interviewed) sample individuals look similar to respondents who cannot be located.

These comments are intended to extend the excellent work presented in the Quality Profiles. The profiles provide details on the measurement of nonsampling error and the results of several experiments to reduce these levels of error. In addition, I hope that as others consider producing quality profiles, these profiles are expanded to cover some of these other issues.

Reference

Duncan, G.J. and Mathiowetz, N.A. A Validation Study of Economic Survey Data, Ann Arbor, MI: The Institute for Social Research, 1985.

Session 2

PARADIGM SHIFTS USING ADMINISTRATIVE RECORDS

PARADIGM SHIFTS: ADMINISTRATIVE RECORDS AND CENSUS-TAKING

Fritz Scheuren¹
Internal Revenue Service

There is a lot in the news lately about problems with the 1990 decennial census in the United States. Many opinions have already been offered about what went wrong and what should be done. Indeed, a paradigm shift may be needed in census-taking.

This brief note talks about the possible role administrative records might play in a new paradigm. To get things started, the word "paradigm" might deserve some elaboration: a paradigm is a way of thinking and then doing; a pattern of belief and behavior; a way of seeing reality and using that sense to accomplish something. Paradigms are common -- the way we get to work would be a humble example. Conventional census-taking, under this definition, could be characterized as a major scientific and technical paradigm.

As long as our paradigms work well for us, we tend not to change them. Occasionally, however, paradigms break down and have to be replaced; e.g., the bridge goes out and we need to find another route to work. As Kuhn pointed out in his seminal book on the structure of scientific revolutions, paradigms break down in science, as well (Kuhn, 1970). Perhaps the most famous example of this is the revolution in the thinking of astronomers that occurred when the Ptolemaic earth-centered view of the universe was replaced by the Copernican view of an earth that revolved, with the other planets, around the sun.

If we look at the problems the U.S. Census Bureau has encountered with the 1990 decennial census, it can easily be argued that one of the major barriers to overcoming these obstacles is the conventional census-taking paradigm. Kish, in a recent paper he has written for Survey Methodology (1990), considers at length some possible alternatives. My objective here will be to focus on two of those areas -- rolling censuses and administrative registers -- and to explore a new paradigm for the U.S. decennial census.

¹By Fritz Scheuren, Director, Statistics of Income Division (R:S), Internal Revenue Service. Based, in part, on a Discussion of "Rolling Samples and Censuses," by Leslie Kish, to appear in the June 1990 issue of Survey Methodology. The views expressed in this paper are those of the author and do not necessarily represent the position of the Internal Revenue Service.

Conventional Census-Taking

Conventional censuses, like those in Canada and the U.S., continue to do many things very well (e.g., Hammond, 1990). Indeed, at present, we have no adequate substitute for them; nonetheless, the need for at least some change seems compelling. Rising costs are a big factor. There have been many improvements in census-taking in this century; still, in both Canada and the U.S., total costs and even costs per person have risen significantly:

- o The 1990 decennial census in the U.S. is budgeted at about \$10 (U.S.) per person. Even adjusting for inflation, this is a four-fold increase over what the per capita expenses were in 1960. Item content differences between the two censuses are small and essentially not a factor in explaining the difference. Both the 1960 and 1990 Census, for example, asked only 7 population questions of everyone (U.S. Bureau of the Census, 1989). The Census long-form sample in 1960 contained 35 questions and was to be completed by 25% of the population. For 1990, the Census long-form sample was given to 16% of U.S. households and had 33 questions.
- o The situation in Canada is similar with regard to the costs of census-taking. For example, the 1991 Canadian Census is budgeted at about \$9.50 (CAN) per person. Like the U.S. Census, there are again just 7, albeit somewhat different, population items that are asked of everyone. Like the 1990 U.S. Census, questions on housing are included for everyone (2 in Canada and 7 in the U.S.). In Canada, a 20% long-form sample will be employed in 1991. The Canadian long-form questionnaire has 45 items for 1991. The 1961 census in Canada was quite different from that planned for 1991 and thus meaningful cost comparisons are hard to make. Nonetheless, looking back 30 years in Canada, the same long-term trend in census-taking costs seems to exist; however, per capita costs have been roughly the same -- even declining slightly -- in the last two or three censuses.

The U.S. Census Bureau has looked at the growing cost of conventional census-taking and concluded that a major change may be needed (Browne, 1989). Labor costs have grown appreciably in recent decades in both Canada and the U.S. Technological improvements have not been great enough to offset these costs, though some, like TIGER (Topologically Integrated Geographic Encoding and Referencing) and CATI (Computer-Assisted Telephone Interviewing), offer promise. Greater attention in the U.S. to improved population coverage is another important factor (Anderson, 1990). The degree of public cooperation in the census also seems to be dropping, at least as reflected by the poorer than

anticipated mail response rate for the 1990 U.S. Census. (It should be noted that this same tendency is not clearly apparent in Canada.)

Increasing cost is not the only major problem facing conventional census-taking. Perhaps of even greater importance is the growing rate of obsolescence of the information collected. The combination of rising costs and growing information obsolescence has had the effect of reducing the benefit/cost ratio for conventional censuses steadily and dramatically.

To obtain more frequent small area data, some countries have introduced quinquennial censuses. For example, in Canada this was first done nationally in 1956. Budget problems led to the 1986 Canadian Census being cancelled and then reinstated. Indeed, it is unclear whether there will be a Canadian Census in 1996. While a quinquennial census was also legislated in the U.S., funds were never made available.

Rolling Censuses

Conventional census-taking, of necessity, must sacrifice both timeliness and item content (on a 100% basis) to achieve complete spatial detail and high population coverage.

One of the alternatives that Kish asks us to look at is a "rolling census." His proposal envisions the sampling of a country over a decade in such a way that every area is eventually covered. In its purest form, space and time become a single dimension and content remains fixed, such that, at decade's end, we have obtained cumulative information on the entire country for a given set of items.

The chief advantage of a rolling census is that it can avoid the problem of information obsolescence at national and major subnational levels. For small geographic areas, though, there would, of course, still be only one observation per decade. Unlike a conventional census, comparisons among small geographic areas would be very difficult to interpret because the data are being collected at different points in time (Fellegi, 1981).

For a rolling census or survey, unit costs could be higher, as Kish notes, than in a more conventional enumeration (indeed, *ceteris paribus*, maybe even higher than the cost of existing survey efforts). In an age of fixed or declining resources, therefore, it might not be possible to do a complete "enumeration" each decade, even if content were significantly scaled back. Rolling samples would seem to have their greatest attractiveness not as a replacement for conventional censuses, but, say, as part of a strategy to link together census-taking with ongoing surveys and

local area population estimates for the intercensal years (Herriot, Bateman and McCarthy, 1989).

Both the United States and Canada employ monthly surveys to estimate the national (and some subnational) labor force characteristics. The Canadian Labor Force Survey (LFS) of 64,500 households covers 0.67% of the total Canadian population each month. "Given the rotation pattern in effect for the LFS, the 0.67% sample per month rolls up into a 6.7% sample of unique households over a 5-year period" (Drew, 1989). In the Canadian context, at least, Kish's proposal may be feasible. A sample survey vehicle could be designed, with some reduction in the month-to-month household overlap, which could achieve many of the benefits he has stated for a rolling sample, while also meeting the information needs currently met by ongoing household surveys (Drew, 1989). This sample would not replace the 100% census count data, itself, but, might be a partial substitute for Canada's 20% long-form census sample.

Because the United States has a population about 10 times larger than Canada, the tradeoffs involving rolling samples and overall country coverage are not as favorable as they are in Canada. The U.S. Current Population Survey (CPS), for instance, at about 60,000 households, covers only .06% of the total U.S. population monthly. Even if cumulated over a whole decade (but, with no change in its rotation pattern), the CPS would cover just roughly 1% of all U.S. households. This does not compare well in size to the overall 16% long-form sample being conducted as part of the 1990 U.S. Census.

To bring the rolling sample population coverage nearer to the 1990 U.S. decennial sample, major changes in the CPS rotation pattern would be needed. Other U.S. Census Bureau surveys might also have to be redesigned if the objective were to achieve even a partial substitute. Despite these changes, moreover, the resulting decade-long sample would still be only a small percent of the total U.S. population -- perhaps, at best, in the 2% to 3% range, assuming resources and other requirements remained essentially fixed.

In both Canada and the U.S., the likely higher unit costs of a rolling sample may need to be addressed by changes in survey procedures: how area segments are listed (Royce and Drew, 1988); how first contact with households is made, etc. Where is it written, for example, that a personal interview contact is needed before using other modes of collection?

It will be no mean challenge to keep effective sample sizes equal for the major level and change components now obtained from ongoing surveys (e.g., Tegels and Cahoon, 1982). Some compromise may be needed, moreover, in the extent to which the basic content of the current long-form census samples can be included. Despite

these challenges, or perhaps because of them, rolling samples deserve continued serious attention and should be the focus of extensive practical experimentation.

Administrative Registers

With the flowering of scientific sample survey methods in the 1940's (Bailar, 1990), the use of administrative records for statistical purposes became relatively less important in many national statistics programs. By the early 1980's, however, at least in the developed countries, the pendulum had begun to swing back. Philip Redfern has been the major chronicler of this phenomenon internationally (Redfern, 1987). While the Danes seem to have gone the farthest (Jensen, 1983 and 1987), major efforts have been made in Canada (e.g., Statistics Canada, 1990) and even some in the U.S. (e.g., Alvey and Kilss, 1990).

A good summary of most of the key barriers to the greater use of administrative registers for census-taking is found in Redfern (1989), including the extensive discussion published with that paper. Perception barriers by the citizens (e.g., in Germany) are mentioned as problems. Psychological barriers by the national statistical service may, however, be of equal or even greater importance. Major scientific "paradigm shifts" generally have this problem (Kuhn, 1970). Certainly, this seemed to be part of the reason for the reception given to the proposal (made by me in 1980) to explore the feasibility of making administrative records an integral part of the U.S. Census of Population. While a sketch of such a proposal was eventually given at the 1982 American Statistical Association meetings (Alvey and Scheuren, 1982), it seems, with a few fairly limited exceptions (e.g., Irwin, 1984; Citro and Cohen, 1985), that serious interest at the Census Bureau has been notably lacking.

Suffice it to say that in the U.S. very little of the needed research has been undertaken. This is true, despite continuing efforts to give the proposal prominence (Jabine and Scheuren, 1985 and 1987) and to get it discussed widely (Butz, 1985). Sadly, therefore, it appears that, in the United States, at least for the year 2000, we should not expect administrative registers to replace censuses.

The 1990 U.S. decennial census could have been used as a proving (or disproving) ground for some of the needed research into administrative record alternatives. Why that didn't happen is a matter that can only be speculated about. A contributing factor, quite possibly, is a case of "paradigm paralysis" (Barker, 1988). The literally decades-long controversy about whether to adjust census "counts" seems to have locked the U.S. Bureau of the Census into what some, at least, would call an increasingly sterile intellectual position (Fienberg, 1990). The viewpoint that they

have adopted makes it very hard for them to see any alternative, like a (partial) administrative record approach, that starts out with the notion that adjustments would be required.

The situation is different in Canada. Since the late 1970's, Statistics Canada has assembled many of the building blocks needed to conduct an administrative record census (e.g., Drew, 1989; Podoluk, 1987; Verma and Raby, 1989). While much remains to be done, such a change could even happen as early as 1996. For example, the coverage of the Canadian tax return system, alone, is quite high and growing. In 1987, for instance, it has been estimated that the coverage was about 94% -- i.e., about 3% less than the 96.8% coverage achieved in the 1986 Canadian Census. By 1991, the tax return coverage, alone, should be up to about 97% or better, with overall administrative record coverage still higher and likely to grow further in the 1990's. (See Table 1 for more details on administrative record coverage in Canada and the U.S.)

TABLE 1 - ESTIMATED ADMINISTRATIVE RECORD COVERAGE
IN CANADA AND THE UNITED STATES: 1987

(Numbers in Millions)

TYPE OF RECORD	Canada		United States	
	Number of Persons	Percent of Total	Number of Persons	Percent of Total
TOTAL POPULATION	26.5	100.0	243.9	100.0
<u>Population Includable in Selected Administra- tive Records:</u>				
Federal Individual Income Tax Records	24.9	94.0	217.5	89.2
Federal Wage Records for Indi- viduals	13.6	51.3	132.9	54.5
Social Security Beneficiaries	3.0	11.3	37.7	15.5
<u>Individuals Not Included Above</u>				
Welfare Records	0.8	3.0	3.1	1.3

Source: Canadian estimates were derived from administrative data at Revenue Canada Taxation and Statistics Canada. U.S. estimates are based on administrative records maintained by the U.S. Department of Treasury and the Department of Health and Human Services.

One concern often raised is that administrative registers, even after they become adequate in quality and coverage, will be

limited to only a few, bare demographic variables: head counts, age, sex and little more. An immediate observation concerning this remark is that conventional censuses do little more than this, themselves, at least for the 100% items. It is also evident that, while the variables on administrative records are not the same as those collected in a traditional census, there is more already available than critics may realize (e.g., Meyer 1990; Alvey and Scheuren, 1982).

More important even than any current item content comparison is the need to emphasize that the proposal to use administrative registers in census-taking does not envision that administrative records have to be used as they are. Administrative records will need to be changed. In my personal opinion, limited optimism about achieving needed changes is justified. However, without a doubt, it is too much to expect of administrative records that they will be able to capture exactly the same concepts now measured in censuses and surveys. Additionally, there almost certainly will need to be special efforts, using existing census-taking techniques, to separately enumerate certain groups. The efforts in the 1990 U.S. Census to count the homeless would be one such example.

Censuses and administrative records each have inherent limitations. Unavoidable conceptual differences will be a major barrier to any shift from one medium to another. Administrative feasibility is another issue; however, some hard-to-duplicate census concepts (e.g., households) may not be as important to the measurement process as formerly.

Shifts in methodology (from conventional census to administrative records) for some uses would potentially be accompanied by a parallel shift in the underlying concepts measured. Some concepts may alter or expand in meaning, including our ability to measure them (e.g., families). We also must ascertain the extent to which respondents answer survey questions the same way they fill out administrative forms that may have real direct impact in their lives.

In recent years, traditional survey methodology has been enhanced by new tools from the field of cognitive psychology. These cognitive research tools could be used to understand any conceptual differences between the meaning of terms when they are used in surveys or drawn from administrative records. We may not have what we think we have anyway (Bates and DeMaio, 1989). In any case, there is already an extensive body of cognitive research that can be drawn on (e.g., Dipbo, 1987; Fienberg and Tanur, 1989; Jobe and Mingay, 1990).

It should also be pointed out that, most likely, administrative registers will not be able to completely meet the demands of modern society for richer sources of statistics. Such

demands, of course, appear to be insatiable. Even if they were not, administrative records will never have the flexibility and responsiveness of surveys. Registers, however, (including partial ones like those that exist in the U.S.) when linked to survey data, can be extremely important as auxiliary variables in making improved direct national survey -- and even subnational survey -- estimates. The U.S. Census Bureau's Survey of Income and Program Participation research on the use of Internal Revenue Service data for improving the precision of national survey estimates is a good recent example (Huggins and Fay, 1988). Indirect (e.g., synthetic) estimates for small areas would still be needed for variables not on the administrative registers (Platek, Rao, Sarndal, and Singh, 1987). The registers, though, might provide a source of valuable symptomatic indicators.

Concluding Observations

The case for considering a "paradigm shift" in census-taking seems compelling, at least in developed countries like Canada and the U.S. The rolling census alternative Kish proposes is probably too expensive to fully implement as a complete substitute for a census. Rolling samples do offer real promise, however, if they can be integrated into the current ongoing survey operations of Canadian and U.S. national statistical programs. Such samples could provide a needed link in addressing small area estimation needs that might otherwise not be met. Less promising, but still possible, is their use as a (partial) substitute for the census long-form samples.

As far as administrative registers are concerned, critics may have been unduly pessimistic. The Canadian situation, however, differs from the United States:

- o In Canada, it is already within the realm of feasibility to combine rolling samples with administrative records as an alternative to conventional census-taking. This is not to say that enormous practical challenges don't remain. The 100% count portion of the Canadian census, though, could be done with administrative records as a starting point, augmented by a large-scale survey to measure and potentially adjust for undercoverage. The Canadian 20% census long-form sample might be, at least partially, replaced by a rolling sample. The content of the Census long-form is considerably richer than that of household surveys, but the content differences could be made up through additional questions "piggy-backing" the on-going surveys at regular intervals. Coverage issues surrounding the use of administrative records could also be addressed directly with rolling samples, especially to calibrate for changes in administrative records between censuses.

- o In the United States, the U.S. Census Bureau has begun to look at alternatives other than conventional census-taking (Bounpane, 1988). Unfortunately, the research needed to look at an administrative register alternative has barely begun. Whether the Census Bureau will find a better approach than the use of administrative records and rolling samples remains to be seen. Whatever other alternatives they study, however, the use of administrative registers as a partial replacement for the conventional 100% counts definitely needs to be considered. A preliminary research agenda updating earlier ideas will appear in Scheuren, Alvey and Kilss, 1990.

Naturally, with such radical proposals, the answer is uncertain. Like Kish, I believe that "the balance of variance components" favors a change from conventional census-taking in most cases. However, as Kish states, "theoretical as well as empirical investigations will be needed to decide matters" (Kish, 1990).

In a change as big as the one proposed here, the "balance" that needs to be struck goes, of course, well beyond looking at variance (and bias) components. One issue that needs to be emphasized more is that some aspects, at least, of the paradigm shifts being considered could go to the heart of the social contract that exists between national statistical agencies and the people that those agencies have a mission to serve. For instance, in the U.S. Constitution, there is a requirement that an "enumeration" of the population take place every ten years. Would the use of administrative records or rolling censuses fit within this "Constitutional paradigm?" Perhaps the starting place is to adopt a broader definition of "enumeration."

Another example where social contract issues arise is the extent to which the greater use of existing (or expanded) administrative data for statistical purposes might be seen as an unwelcome increase in the intrusiveness of the State into the private lives of its citizens (Grace, 1989). As legitimate as concerns about "intrusiveness" might be, though, there is no evidence in a North American context, at least, that they pose an insurmountable barrier. On the contrary, there have been virtually no adverse public reactions to past U.S. additions to administrative records for statistical purposes (e.g., of residential address information in 1972, 1974 and 1980 tax returns). To my knowledge the issue, so far, has not come up directly yet in Canada, at least at the Federal level.

In summary, to make these kinds of changes there is the need for a lot more scientific research. Studying the implementation technologies will be an even bigger job. Finally, the issues go beyond our profession and may well be settled in other arenas. Wherever they are decided, it is incumbent on us, as statisticians,

to frame the debate in terms of feasible options. Hopefully, exchanges such as ours today will help lead the way along that path.

References

Alvey, Wendy and Kilss, Beth (eds.) (1990). Statistics of Income and Related Administrative Record Research, U.S. Department of the Treasury, Internal Revenue Service. See also Kilss, Beth and Alvey, Wendy (eds.) (1984). Statistical Uses of Administrative Records: Recent Research and Present Prospects, vols. 1 and 2, U.S. Department of the Treasury, Internal Revenue Service.

Alvey, Wendy and Scheuren, Fritz (1982). "Background for an Administrative Record Census," 1982 American Statistical Association Proceedings, Social Statistics Section, pp. 137-146.

Anderson, Margo (1990). "'According to Their Respective Numbers ...' for the Twenty-First Time," Chance, vol. 3, no. 1, pp. 12-18.

Bailar, Barbara (1990). "Contributions to Statistical Methodology from the Federal Government," Survey Methodology, vol. 16, no. 1, Statistics Canada.

Barker, Joel Arthur (1988). Discovering the Future: The Business of Paradigms, Institute for Information Studies.

Bates, Nancy A. and Demaio, Theresa J. (1989). "Using Cognitive Research Methods to Improve the Design of the Decennial Census Form," Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, 267-285.

Bounpane, Peter (1988). "A Sample Census: A Valid Alternative to a Complete Count Census?" 46th Session of the International Statistical Institute.

Browne, David (1989). "U.S. Bureau of the Census: Facing the Future Labor Shortage," Asian and Pacific Population Forum, vol. 3, no. 4.

Butz, William (1985). "Comment: The Future of Administrative Records in the Census Bureau's Demographic Activities," Journal of Business and Economic Statistics, vol. 3, no. 4, pp. 393-395.

Citro, Connie and Cohen, Michael L., eds. (1985). The Bicentennial Census: New Directions for Methodology in 1990. National Academy Press, Wash., DC.

Dippo, Cathy (1987). "A Review of Statistical Research at the U.S. Bureau of Labor Statistics," Journal of Official Statistics, vol. 3, no. 3, pp. 289-297.

Drew, J. Douglas (1989). "Address Register Development and its Possible Future Role in Integration of Census, Survey and Administrative Data," A paper presented at the U.S. Bureau of the Census/Statistics Canada Interchange. (Unpublished)

Fellegi, I.P. (1981). "Comments," Discussion of a paper by Leslie Kish entitled "Population Counts from Cumulated Samples," Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau, An Analysis, Review and Response, Congressional Research Service, the Library of Congress.

Fienberg, Stephen (1990). "An Adjusted Census in 1990? An Interim Report," Chance, vol 3, no. 1, pp. 19-21.

Fienberg, Stephen and Tanur J. (1989). "Combining Cognitive and Statistical Approaches to Survey Design," Science, 243, pp. 1017-1022.

Grace, John W. (1989). "The Use of Administrative Records for Social Research," Statistics Canada Workshop, December 12, 1989, Ottawa, Ontario.

Hammond, Robert B. (1990). "The 1990 Decennial Census: An Overview," Conference Proceedings, Advanced Computing for the Social Sciences, sponsored by the Oak Ridge National Laboratory and the U.S. Bureau of the Census, April 10-12, 1990, Williamsburg, Virginia.

Herriot, Roger; Bateman, David V.; and McCarthy, William F. (1989). "The Decade Census Program -- New Approach for Meeting the Nation's Needs for Sub-National Data," to appear in American Statistical Association Proceedings, Social Statistics Section.

Huggins, Vicki and Fay, Robert (1988). "Use of Administrative Data in SIPP Longitudinal Estimation," American Statistical Association Proceedings, Section on Survey Research Methods.

Irwin, Richard (1984). "Feasibility of an Administrative Records Census in 1990," Special Report on the Use of Administrative Records, Committee on the Use of Administrative Records in the 1990 Census, unpublished Census Bureau report.

Jabine, Thomas B. and Scheuren, Fritz (1985). "Goals for Statistical Uses of Administrative Records: The Next Ten Years," Journal of Business and Economic Statistics, vol. 3, no. 5, pp. 380-391.

Jabine, Thomas B. and Scheuren, Fritz (1987). "Statistical Uses of Administrative Records in the United States: Where Are We and Where Are We Going?" Proceedings of an International Symposium on Statistical Uses of Administrative Data, J.W. Coombs and M.P. Singh (eds.), Statistics Canada, December 1988, Ottawa, pp. 43-72.

Jensen, Poul (1983). "Towards a Register-Based Statistical System -- Some Danish Experiences," Statistical Journal of the United Nations Economic Commission for Europe, vol. 1, no. 3, pp. 341-365.

Jensen, Poul (1987). "The Quality of Administrative Data from a Statistical Point of View: Some Danish Experience and Consideration," Proceedings of an International Symposium on Statistical Uses of Administrative Data, J.W. Coombs and M.P. Singh (eds.) Statistics Canada, Ottawa.

Jobe, Jared B. and Mingay, David J. (1990). "Cognition and Survey Measurement: History and Overview," Applied Cognitive Psychology, in press.

Kish, Leslie (1990). "Rolling Samples and Censuses," Survey Methodology, in press.

Kuhn, Thomas S. (1970). The Structure of Scientific Revolutions, Second Edition, Enlarged, The University of Chicago Press, Chicago.

Meyer, Bruce (1990). "The Tax System: Comparisons of Demographic, Labour Force and Income Results for Individuals and Families," Small Area and Administrative Data Division, Statistics Canada.

Platek, R.; Rao, J.N.K.; Sarndal, E.E.; and Singh, M.P. (1987). Small Area Statistics, New York: Wiley-Interscience.

Podoluk, J. (1987). "Administrative Data as Alternative Sources to Census Data," Proceedings of an International Symposium on Statistical Uses of Administrative Data; J.W. Coombs and M.P. Singh (eds.), Statistics Canada, December 1988, Ottawa, pp. 273-290.

Redfern, Phillip (1987). "A Study of the Future of the Census of Population: Alternative Approaches," Eurostat Theme 3 Series C, Luxembourg: Office for Official Publications of the European Communities.

Redfern, Phillip (1989). "Population Registers: Some Administrative and Statistical Pros and Cons," The Journal of the Royal Statistical Society, Series A (Statistics in Society), vol. 152, pt. 1, pp. 1-41.

Royce, Don and Drew, J. Douglas (1988). "Address Register Research: Current Status and Future Plans," 1991 Research and Testing Project, 1991 Census, Statistics Canada, Ottawa.

Scheuren, Fritz; Alvey, Wendy and Kilss, Beth (1990). "Paradigm Shifts: The Integration of Administrative Records and Surveys," a paper delivered at the 151st Annual Meeting of the American Statistical Association, August 7, 1990, in Anaheim, CA.

Statistics Canada (1990). "Research Papers and Reports," Bibliography, Small Area and Administrative Data Division, Ottawa, Ontario. (unpublished)

Tegels, Robert and Cahoon, Lawrence S. (1982). "The Redesign of the Current Population Survey: The Investigation into Alternate Rotation Plans," Proceedings of the American Statistical Association, Survey Research Methods Section.

U.S. Bureau of the Census (1989). 200 Years of U.S. Census Taking: Population and Housing Questions, 1790-1990, Superintendent of Documents, U.S. Government Printing Office, Washington, DC.

Verma, Ravi B.P. and Raby, Ronald (1989). "The Use of Administrative Records for Estimating Population in Canada," Survey Methodology, vol. 15, no. 2, pp. 261-270.

AN ADMINISTRATIVE RECORD PARADIGM:
A CANADIAN EXPERIENCE²

John Leyes
Statistics Canada

1.0. Introduction

In 1979, Statistics Canada began a formal review of the potential of using administrative records for social statistical applications for small area data (Statistics Canada, 1979). Based on this review, it was concluded that the highest coverage of the population and the greatest potential for social administrative data would arise through the use of the personal income tax records. With few exceptions, then, this paper considers data derived from the personal income tax file in Canada.

The Canadian tax system differs from the U.S. tax system. For example, in Canada, there is no joint filing; and the tax system is used as an instrument to provide benefits to persons and families with low incomes. The personal income tax return is known as the T1. The T1 serves a purpose similar to the IRS' Form 1040.

In its earliest days, Statistics Canada's work with the personal income tax file was subject to a number of expected a priori shortcomings. These shortcomings represented an administrative records paradigm (or rules of the game). The shortcomings included the following:

- Population Coverage. The income tax system is based on individuals only. Since only 60% of Canadians were filing tax returns in the mid-1970's, coverage was deemed inadequate for social statistical applications.
- Population Coverage Bias. The age profile of taxfilers differed from the age profile of the population. This was judged to be an unacceptable bias.
- Income Coverage. Not all income received by Canadians is taxable; hence, the income coverage of the T1 was considered incomplete.
- Income Distribution Coverage. Since both the elderly and the young frequently have low incomes and do not file tax returns, data from the tax file would be inadequate for public policy purposes directed at these target groups.

² This is a summary of a longer paper that was prepared for the Seminar on Quality of Federal Data.

- Dimensionality of Variables. Since any single administrative record has a specific and narrow application in program administration, the range of data variables were also judged inadequate as a source of social data.
- Concepts and Definitions. The concepts and definitions used in household surveys and censuses of population can only be approximated through the use of an annual tax file.

Each of the above represented a limitation or shortcoming for data derived from administrative records in general, the T1 in particular. In spite of these shortcomings, the work began, and this paper is directed at a few findings that resulted from work in Canada with the personal income tax records in the development of family data since 1984. Perhaps this paper may even indicate some of the potential of using the T1 as a source of small area data in post-censal periods in Canada.³

2.0. The Development of Taxfiler Family Data

The taxfiler family concept has been designed to emulate the census family concept. A census family:

"[r]efers to a husband and a wife (with or without children who have never married, regardless of age), or a lone parent of any marital status, with one or more children who have never married, regardless of age, living in the same dwelling. For census purposes, persons living in a common-law type of arrangement are considered as now married, regardless of their legal marital status; they accordingly appear as a husband-wife family in census family tables." (Statistics Canada, 1982, p. 29)

This concept is suitable for household collection methods since respondents are asked to report on the relationships between all residents of a dwelling. With administrative records, secondary information such as reported marital status, value of exemptions/tax credits, ages of taxfilers, addresses, child care expenses, and so forth, are used for forming families.

It has not, therefore, been possible to emulate the exact census family concept. The major sources of difficulty arise with older children (whether they have ever been married or not when they reside with their parents) and with common law couples. In

³ A recent bibliography of the staff papers and reports prepared on the use of administrative records for social data in the Small Area and Administrative Data Division was recently completed. (Statistics Canada, 1990)

general, the census family concept works reasonably well for families with dependent children, and some success has been achieved in estimating single parent and common law families, as can be seen in Table 1.

In 1984, Statistics Canada began estimating families from the individual taxfiler (T1) data. The creation of families from the T1 is based on a six-step process:

- i. Taxfilers, reporting the Social Insurance Numbers (SIN) of their spouses, are matched to form husband-wife families;
- ii. Other husband-wife families are formed from taxfilers who declare themselves married but do not report spousal SINS;
- iii. Non-dependent filing children⁴ who reside with their parents are matched to their parents;
- iv. There is an intermediate step to unduplicate records, to identify one-filer husband-wife family units, to assign a unique postal code to family members, and to assign a family composition type to each family unit;
- v. Common law spouses are matched from the pool of individuals classed as single parent families and non-family persons; and
- vi. In Step 6 non-filing family members are imputed.

With this brief introduction and description of the taxfiler family data, it is now possible consider some data findings.

3.0. The Coverage Shortcomings, Some Empirical Findings

3.1. Population Coverage Comparison: 1985 Taxfiler Family File (T1FF) to 1986 Census of Population⁵

The taxfiler family data have been placed into four classifications: husband-wife families, single parent families, common

⁴ Unmarried persons who (a) declare themselves to be single, (b) are under the age of 30, (c) reside with their parents and (d) file a tax return are defined to be "filing children".

⁵ To minimize the T1FF data processing costs, most of the T1FF data in this paper are based on samples.

law families, and non-family persons. The data in Table 1 reflect the first three of these classifications (common law families are noted twice, once as husband-wife families, and then separately).

In creating the taxfiler family (T1FF) data, a record is created for each family member and for each non-family person. Thus, there is a record for a taxfiler and a record for each person that is imputed. Line three of Table 1, therefore, is an estimate of the T1FF population that can be identified through the tax system.

TABLE 1.
Summary of Comparisons Between Tax Family & Census Data, 1982 - 1988.

LINE NO.	TAX YEARS						
	1982	1983	1984	1985	1986	1987	1988
NUMBER OF TAXFILERS							
1. Taxfilers (000's)	15,166	15,243	15,467	15,526	15,971	16,687	17,255
2. % increase		0.5	1.5	0.4	2.9	4.5	3.4
3. TOTAL TAX POP. (000's)	23,628	23,725	23,736	23,839	24,016	24,838	n/a
4. % increase		0.4	0.0	0.4	0.7	3.4	
5. Pop. Est. (,000)	24,787	24,978	25,165	25,353	25,625	25,923	n/a
6. Coverage (Tax/Census)	95.3	95.0	94.3	94.0	93.7	95.8	
HUSBAND-WIFE FAMILIES							
7. Tax (000's)	5,510	5,524	5,570	5,528	5,592	5,753	5,882
8. Census (000's)	5,722	5,773	5,824	5,875	5,932	5,987	n/a
9. Coverage (Tax/Census)	96.3	95.7	95.6	94.1	94.3	96.1	
SINGLE PARENT FAMILIES							
10. Tax (000's)	830	873	894	934	940	965	n/a
11. Census (000's)	768	796	824	852	882	912	
12. Coverage (Tax/Census)	108.1	109.7	108.5	109.6	106.5	105.7	
***** COMMON LAW FAMILIES ARE INCLUDED IN LINE 7 *****							
COMMON LAW FAMILIES							
13. Tax (000's)	205	207	242	221	281	336	367
14. % increase		0.8	16.8	-8.5	27.3	19.4	9.3
15. Census (000's)				487			
16. Coverage (Tax/Census)				45.4			

Notes:

- Line 5 is based on population estimates only.
- The 1982, 1983 and 1985 single parent tax family data are based on a 5% sample.

SOURCES: 1985 T1FF. Preliminary, unpublished, 10% sample.

Population Estimates. Catalogue No 91-204. June, 1988

There are several highlights in Table 1:

- The T1FF population has varied between 93.7 and 95.8 percent during the 1982 and 1987 period (line 6 of Table 1);⁶
- The number of total tax family records (taxfilers plus imputed) increased at a slightly lesser rate than the number of taxfilers alone (i.e. lines 2 and 4);
- Coverage of husband-wife families was slightly higher than coverage of total family records (i.e., lines 6 and 9); and
- Single parent overcoverage decreased for 1986 and 1987.

3.2. Population Coverage Bias, 1985 T1FF to 1986 Census

In Table 2, some broad age range comparisons have been included. The first age range is, perhaps, a bit unusual since it includes the population 29 years of age and below. This age range resulted from an arbitrary decision, namely, that the maximum age of a matched filing child could be 29. Furthermore, for imputed, non-filing dependent children, there is limited age information and no gender information. Thus, children, whether imputed as dependents or identified as taxfilers who reside with their parents, have been placed into one age range.⁷

In reviewing column 4 of Table 2 (i.e., % ratio), it can be noted that the coverage of the T1FF to the 1986 Census was approximately 90% or higher for age ranges under 60.⁸ The T1FF coverage of the 1986 Census population declined more rapidly for the population 65+.

⁶ In processing the 1986 tax file, a somewhat earlier file was used than in other years. As a result, the coverage was lower than in other years. Had this not occurred, the coverage in 1986 would have been higher than 93.7%.

⁷ The T1 does contain some information on dependent children, namely, relationship to taxfiler and birthdate. This information is not, however, captured.

⁸ The taxfiling rate for the 65+ population increased from 60% in 1985 to 75% in 1987.

TABLE 2.
Population Comparison by Age Ranges,
1985 T1FF to 1986 Census.

AGE GROUP	1985 T1FF (000'S)	1986 CENSUS (000'S)	% RATIO (T1FF/CENSUS)	1985 T1FF IMPUTATIONS	% of T1FF RECORDS IMPUTED
0 - 29	12,044.2	11,911.7	101.1	6,920.8	57.5
30 - 34	2,083.5	2,185.6	95.3	113.6	5.5
35 - 39	1,921.1	2,026.2	94.8	150.7	7.8
40 - 44	1,500.6	1,614.7	92.9	135.3	9.0
45 - 49	1,224.2	1,315.9	93.0	123.5	10.1
50 - 54	1,137.4	1,229.3	92.5	125.9	11.1
55 - 59	1,079.6	1,203.2	89.7	162.6	15.1
60 - 64	982.1	1,125.1	87.3	192.5	19.6
65 - 69	728.3	911.8	79.9	148.0	20.3
70 - 74	528.0	738.3	71.5	97.4	18.4
75 +	577.8	1,047.5	55.2	83.3	14.4
TOTAL	23,806.7	25,309.3	94.1	8,253.7	34.7

SOURCES 1985 T1FF Preliminary, unpublished and 10% sample data.
1986 Census. Unpublished tabulations.

3.3. Coverage of Aggregate Sources of Income

In conducting the 1986 Census of Population, sources of income data were collected for the 1985 calendar year. Table 3 contains a sources of income comparison between the 1985 T1FF and the 1986 Census.

For both data sources, the largest component of income was wages and salaries. The T1FF estimate was 96% of the Census estimate. In the government transfers section of Table 3, considerable variability existed, primarily because some transfer payments were either not subject to taxation or were received by individuals with low incomes who did not file a T1.

3.4. Income Distribution Coverage

Table 4 includes a time series comparison of median incomes between the T1FF and the Survey of Consumer Finances (SCF)⁹ for the period 1982-87. The T1FF medians were lower for all years. Moreover, the medians were about 95% for the first four years. In the fifth and sixth years, the medians declined to about 92%. This decline can be partly attributed to the introduction of a refundable Federal Sales Tax Credit. This credit was available to

⁹ The SCF is an annual supplement to the Canadian Labour Force Survey. The SCF is similar to the March supplement to the Current Population Survey (CPS) in the United States.

individuals and families with low incomes, some of whom may only file a tax return to obtain this credit.

TABLE 3.
Sources of Income Comparison (\$000,000,000's)
1985 T1FF to 1986 Census

SOURCES OF INCOME	1985 T1FF	1986 CENSUS	% RATIO (T1FF/CENSUS)
Wages & salaries	217.9	227.1	95.9
Self-employment	15.1	17.9	84.4
Investment	24.9	20.5	121.5
Pension	10.6	8.7	121.8
SUBTOTAL	268.5	274.2	97.9
GOVERNMENT TRANSFERS			
	9.9	16.1	61.5
OAS/GIS/SPA + C/QPP	9.4	7.2	130.6
Unemployment Ins.			
Family Allowance +	3.9	4.0	97.5
Child Tax Credit	0.0	7.3	0.0
Other Gov Transfers			
SUBTOTAL	23.2	34.6	67.1
Other	2.3	2.7	85.2
TOTAL	294.0	311.5	94.4

SOURCES: 1985 T1FF. Preliminary, unpublished and 10% sample data, plus unpublished Child Tax Credit data from the Economic Dependency Profile, Small Area and Administrative Data Division, Statistics Canada. 1986 Census, Catalogue No. 93-114, Table 2.

Since it is generally assumed that taxfilers have higher incomes than non-taxfilers, one would expect the SCF to have lower medians since some respondents would have low incomes and not file tax returns. Clearly, these findings are inconsistent with such an expectation.

3.5. Dimensionality of Variables

Since any single administrative record (for example, the T1) has a specific and narrow application in program administration, the range of data variables might be judged inadequate as a source of social data. Although the T1FF data are oriented to the income tax system, Table 5 indicates (mainly by reference to the footnotes) that some comparability in the variables existed between the 1985 T1FF and the 1986 Census.¹⁰

¹⁰ This table was adapted from Vigder and Leyes (1989).

TABLE 4.

Income Comparison for Families, 1982-87:
TIFF to Survey of Consumer Finances

YEAR	MEDIAN INCOME, FAMILIES		% RATIO TIFF/SCF
	TIFF	SCF	
1982	28,154	29,537	95.3
1983	28,806	30,419	94.7
1984	30,603	32,079	95.4
1985	32,140	33,950	94.7
1986	FEDERAL	SALES TAX	CREDIT INTRODUCED
1986	33,135	36,019	92.0
1987	35,279	38,059	92.7

SOURCES: 1982-87 TIFF. Preliminary, unpublished and 5% sample data. SCF data, Catalogue No. 13-208.

From Table 5, it is clear that the TIFF data lack the richness of the census data. TIFF has a low coverage of non-taxable sources of income and a low coverage of those taxable sources of income received by low income persons who do not file tax returns.

4.0. Major Directions for 1989+

Two new initiatives have been started.

- The development of a pilot Longitudinal Administrative Database (LAD) to enable research studies of poverty/welfare/income dynamics in Canada for the period 1982-86. The LAD was designed as a 10% sample to parallel the Panel Survey of Income Dynamics (PSID) that was begun by the Survey Research Center, University of Michigan about 20 years ago. (Duncan, 1984)
- The development of an Administrative Record Consolidation File (ARC) through the linking of multiple records on a sample basis for the purpose of (a) improving the coverage of the population and (b) improving some of the variables on the taxfile.

5.0. Summary and Related Observations

The TIFF data possess some positive characteristics. The data are annual and small area estimates can be produced. Furthermore, if 95% is high coverage, the comparisons in this paper have indicated a fairly high coverage of the population by the TIFF.

One potential benefit of administrative data seems to lie in the domain of longitudinal databases. While longitudinal surveys can only be created in the future, based on current decisions and funding, longitudinal administrative databases can be created retrospectively, based on current decisions and funding. For example, a decision was made in late 1988 to begin creating a longitudinal database for the period 1982 to 1986. While the database has not yet been completed, early indications are that the database will be a source of useful information for the development of social policy and for the analysis of income dynamics.

To conclude, this paper has been prepared to illustrate some findings that may not be widely known. In preparing this incomplete report on an evolving new paradigm in Canada, it is hoped that members of the research and statistical community will provide comments and insights that will improve and stimulate the continued evolution of this work.

TABLE 5
Comparison of 1986 Census Content to
1985 Taxfiler Family (TIFF) Data

	1986 CENSUS	1985 TIFF
100% DATA COMPARIAONS		
Exact Location	x	
Mailing Address		x
Owned or Rented	x	
Number of:		
Husband-Wife Families	x	x
Lone Parent Families	x	x
Common Law Families	x	x
Non-Family Persons	x	x
Number of Persons	x	x
Relationship to Person 1	x	x (1)
Gender	x	x (2)
Birthdate	x	x (2)
Marital Status	x	x (2)
Mother Tongue	x	x (3)
Aboriginal Status	x	
20% SAMPLE DATA		
Housing	x	
Education	x	x (4)
Place of Birth	x	
Citizenship	x	
Ethnic Group	x	
Language Spoken at Home	x	
100% Versus SAMPLE DATA		
Income	Sample	100 %
Labour Market Activity	Sample	x (5)
Occupation	Sample	x (6)
Industry	Sample	x (7)
Residence 5 Years Ago	Sample	x (8)

Notes:

- (1) For children dependents, this information is not captured.
- (2) This is available for taxfilers only.
- (3) The language (French/English) of the form is captured.
- (4) Place of birth is recorded on the Social Insurance Number (SIN) Master File.
- (5) The presence/absence of employment income is on the TIFF for all taxfilers.
- (6) A crude occupation question is requested of all taxfilers.
- (7) A link with the T4 and Business Register can yield industry data for all taxfilers.
- (8) Although not currently available, this is possible with TIFF with a five-year link of TIFF records.

References

Duncan, Greg J. (1984) Years of Poverty, Years of Plenty. Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan. 1984.

Statistics Canada. (1979) Final Report of the Task Force on Administrative Data Development. Unpublished. Ottawa, Canada. March 16, 1979.

Statistics Canada. (1982) 1981 Census Dictionary. Statistics Canada, 1981 Census of Canada. Catalogue Number 99-901. Ottawa, Canada. May 1982.

Statistics Canada. (1988) Postcensal Estimates of Families, Canada, Provinces and Territories. Catalogue Number 91-204. Ottawa, Canada. 1988.

Statistics Canada. (1989) Total Income: Individuals. 1986 Census of Canada. Catalogue Number 93-114. Ottawa, Canada. March 1989.

Statistics Canada. (1990) Bibliography: Research Papers and Reports, Small Area and Administrative Data Division. Small Area and Administrative Data Division, Statistics Canada. Ottawa, Canada. Unpublished. February 22, 1990.

Statistics Canada. (Annual) Family Incomes: Census Families. Survey of Consumer Finances. Catalogue No. 13-208 (Annual). Ottawa, Canada. 1989.

Vigder, Michele and John Leyes. (1989) Administrative Data to Mid-Decade Census Data Comparison. Small Area and Administrative Data Division, Statistics Canada. Ottawa, Canada. Unpublished. January 31, 1989.

DISCUSSION

Gerald Gates
U.S. Bureau of the Census¹¹

The theme of these papers by Fritz Scheuren (IRS) and John Leyes (Statistics Canada) is shifting the paradigm of census taking to allow for more frequent detailed information for small geographic areas at reasonable costs. Scheuren points to two weaknesses in the U.S. census taking process -- 1) the increasing costs of enumeration and 2) the increasing obsolescence of the information between censuses. He discusses new paradigms that have been proposed by Kish and others which employ rolling samples and other techniques to obtain more frequent small area data. His primary focus, however, is on administrative registers that could be modified to serve a census function as well as their intended administrative uses. His intent is to frame the debate for feasible options that will lead to a lot more scientific research on this topic.

Leyes addresses the census paradigm in terms of research undertaken by Statistics Canada using administrative records. Primarily, he describes the development of a family tax file representing approximately 95% of the census in terms of population covered. He looks at coverage of this file in comparison with the census; with surveys conducted by Statistics Canada; and with administrative data maintained by other agencies. Finally, he describes a project to develop a linked administrative file that would allow Statistics Canada to estimate the characteristics of the population missed in the family tax file. The work Leyes describes has implications for shifting the census paradigm to address cost, accuracy, and timeliness issues.

Turning first to the Leyes' paper, I have a few specific reactions to the role Statistics Canada plays with Revenue Canada and with the content and coverage of the family tax file. The family tax file could only have been created with a great deal of cooperation from Revenue Canada. The Canadian tax form contains demographic characteristics such as age, sex and marital status that have no practical tax program application. In addition, all information from the tax return is available to Statistics Canada -- this is not the case in the U.S. Another major difference between the two countries is the negative income tax provisions in Canada which increases coverage of the tax file

¹¹These remarks are attributable to the author and do not necessarily represent the views of the Census Bureau.

(from 89% in the U.S. to 96% in Canada). Despite these advantages, the Canadian tax form, like the IRS Form 1040, contains mailing address rather than physical address.

I also found the Canadian work on record linkages to be quite impressive, especially as it relates to creating retrospective longitudinal databases to deal with emerging issues (e.g., income and health care issues relating to the elderly). Also, these linkages permit, as Leyes states, adjustment of the family tax file for undercoverage. This feature allows Statistics Canada to use the family tax file as an independent source for producing population estimates between census years. Since these linkages are only done on a sample basis due to privacy concerns, their utility is diminished somewhat. In the current U.S. situation, the reduced coverage and content of the Form 1040 file makes 100% record linkages critical, while similar privacy concerns need to be addressed. (I should add that Form W2 earnings records could improve the coverage possible with only 1040 tax returns, but this will continue to miss nonworkers and omit some of the detail available on the 1040.)

Scheuren's paper raises some important issues regarding the need for research on alternatives to the traditional once-a-decade enumeration. I complement Fritz on his persistence over the years to explore traditional census alternatives. His current paper addresses the need for a census alternative to deal with "problems" facing the 1990 census in terms of costs (low mail response rates) and increasing data obsolescence. Although administrative records remain his primary focus, Fritz sees a need for research in other areas, such as rolling samples. He believes that rolling samples offer real promise if they can be integrated into current ongoing survey operations. Although the "rotating" sample techniques have been proposed for 2000 census planning (Herriot, Bateman, McCarthy, 1989), little research has been done and we have no plans to incorporate this technique into the current surveys. There are several reasons for this which reflect the different goals of current surveys and intercensal estimates:

- o a rolling design will create inefficiencies because of increased interviewer travel (and reduced workloads) which will come from abandoning Primary Sampling Units (PSUs) in favor of more geographically disperse samples;
- o survey procedures that, as a cost saving feature, incorporate an alternative to the traditional first time personal visit, could result in lower response rates (telephone) or delays in the interviewing process (mail);
- o for surveys such as SIPP, the sample may be too small to spread out geographically;

- o sponsors may not want long form questions added to their questionnaires nor want intensive sample in areas with small population.

The second major point I would like to address regards, as Fritz puts it, our "missed opportunity" to use the 1990 census as a proving ground for the use of administrative records in the census process. The 1990 census Research and Experimentation (REX) program considered many applications for administrative records including all uses made in 1980 plus an administrative records census and a coverage improvement program designed to enumerate parolees and probationers through state administrative records. All of these uses were abandoned because of resources available and the expected minimal improvements given the costs. (The parolee/probationer operation was accomplished by parole officers who distributed and collected questionnaires from persons in their charge.) An additional use, which was tested on a small scale as part of the 1988 dress rehearsal, involved supplementing the Post Enumeration Survey (PES) with names obtained from administrative records in order to improve the PES as a coverage measurement tool. (Wolfgang, 1989) An evaluation of this test, which will be released shortly, may encourage further research in this area.

Several administrative records uses that were adopted in the 1990 census include:

- o use of local lists of shelters and street locations to assist in enumerating the homeless;
- o use of vendor lists for developing the mail register;
- o macro-level consistency checks for content evaluation;
- o encouraged use by local jurisdictions as a way of improving outreach activities.

Like Fritz, I believe that more extensive use of administrative records, in a productive way, will require changing administrative records. But, it will take more than that. It will take institutional changes in the way administrative agencies view their role in the census statistical process.

By way of tying this challenge to the future research activities of the Census Bureau, allow me to expand slightly on Fritz' paradigm analogy and relate it to the environment in which we operate. Both Leyes and Scheuren see administrative records playing a key role in shifting the "census" paradigm. Under this assumption, I suggest that, rather than a single census paradigm, there are actually three interrelated paradigms that require equal consideration. These are the once-a-decade enumeration, intercensal population estimates, and administrative records information systems.

Before we consider approaches to shifting these three paradigms we need to think about the role the public and bureaucracies may play. We need to consider the social contracts that exist between the government and the American people. The statistical agency has a specific obligation to census respondents to ensure privacy (confidentiality) and reduce burden to the extent possible. In addition, the statistical agency must fulfill its obligation to the American taxpayer to use its resources in the most efficient manner in providing the information needed by society. Balancing these tradeoffs will determine which direction the paradigm shift takes.

Shifting the administrative records paradigm also requires a new partnership between Federal agencies and, possibly, between Federal agencies and the states. The administrative agency must accept new unrelated tasks that are not part of its primary mission. Traditionally, agencies avoid taking on tasks that differ significantly from those that are at the heart of the organization's mission. (Wilson, 1989) Even within an administrative agency, the statistical functions often take a back seat to administrative functions. Despite laws and additional funds that reflect these new tasks, when push comes to shove, the primary mission (in the case of the IRS, collecting taxes) will most likely win out.

A census reliance on administrative records requires a commitment by the administrative agency to the census function which heretofore has not existed. Where information is lacking, such as physical address and household relationships, change must be encouraged. Where change in the administrative process could negatively affect the census use, accommodations must be made. In the past, changes have occurred but they have not always been anticipated or beneficial. For example,

- o Physical location information was added previously to the Form 1040 by the Census Bureau for the General Revenue Sharing Program.
- o The 1986 Tax Reform Act required the IRS to collect SSNs for children (a plus) but eliminated the personal exemption for persons 65 or older (a negative).
- o The SSA recently introduced a program of assigning SSNs to infants at birth using state birth records. Despite Census Bureau objections and concerns of its own statistical office, SSA did not require that race of child (or mother) be part of the application process.

If we assume that planning for paradigm shifts is good -- which I think we must -- then we need to consider, as Fritz suggests, which options are feasible. First let me discuss options as they relate to the traditional census. The basic Constitutional

requirement for apportionment requires an actual enumeration every ten years. Seven items are requested from each resident to provide: 1) the basis for the apportionment of Congress; 2) a sampling frame for use in the next decade; and 3) a base for developing intercensal estimates. To obtain this information from administrative records (i.e. an administrative records census), may require a Constitutional amendment in addition to changes to the way administrative agencies do their jobs. Research on this aspect should concentrate on the most useful sources of information with the least amount of change required.

A second component of the census consists of the housing questions asked of every household. The Census Bureau is exploring the possibility of obtaining this information in future censuses from records of city or county tax offices, assessors offices, or recorders offices. Such an option has the potential to reduce burden and costs of census taking while offering comprehensive coverage of the nation's housing. One of the key requirements for such an operation would be fostering interest in the local jurisdictions to change/standardize their information systems to maintain the items needed for the census. This could be done by promoting the changes as an improvement to existing administrative systems and as a rich source of data for administering housing related programs. In this way, we win acceptance for the changes needed for statistical purposes through the administrative benefits they provide.

The final component of the census is the long form sample questions. This component provides a source of detailed information for small geographic areas -- but only once a decade. As Scheuren suggests, these data could come from a rolling census design in the event that the basic census (count) is done through administrative records, but there are many problems as I have noted.

The intercensal estimates paradigm is certainly tied to the census paradigm and any change to the census will most likely necessitate changing the way we do intercensal estimates. The current population estimates program was a byproduct of the General Revenue Sharing Program. We will evaluate alternative designs in the years ahead to see if the current program is meeting the needs of users. The work of Statistics Canada on developing family tax files definitely needs to be considered. In addition, recently proposed legislation would put greater reliance on currently available population estimates for funds allocation formulas which will in turn put pressure on the Census Bureau to expand the utility of these estimates. A possible alternative which is being given some consideration by the Census Bureau would involve conducting a large sample survey at mid-decade and modeling the results to administrative records linked to TIGER geography. The administrative file could be constructed by linking the tax returns obtained by Census with the social security number applicant file

to be obtained from the Social Security Administration (assuming we can address the privacy issues).

In conclusion, greater reliance on administrative records in the census process needs public acceptance and a commitment from all those affected to make it work. Perhaps the increasing costs and respondent burden involved in traditional census taking will encourage this change. Scheuren and Leyes have shown us some options. We will need to explore these and others -- and fund the necessary research -- so that, as we move into the 21st Century, we are able to avoid the pitfalls and take advantage of the opportunities that lie ahead.

References

Herriot, R.; Bateman, D.; and McCarthy, W. (1989). "The Decade Census Program -- New Approach for Meeting the Nation's Needs for Sub-National Data," to appear in American Statistical Association Proceedings, Social Statistics Section.

Wilson, J. (1989). Bureaucracy -- What Government Agencies Do and Why They Do It; New York: Basic Books, Inc., p. 190.

Wolfgang, G. (1989). "Using Administrative Lists to Supplement Coverage in Hard-to-count Areas of the Post-Enumeration Survey for the 1988 Census of St. Louis." Proceedings of the Section on Survey Research Methods, American Statistical Association.

DISCUSSION

Edward J. Spar
Market Statistics

The Scheuren paper, is provocative and challenging. At the same time, some of the ideas presented here should be challenged back. For example, Scheuren mentions how expensive the decennial census has become - \$10 per capita. But based upon what is this expensive, in other words, as compared to what? If each individual, has to spend about \$1 a year for the decennial census, is this still considered to be too expensive? Maybe we should have a check off box on the 1040 form for those who wish to contribute a dollar to the census instead of presidential election campaigns. Money better spent.

Scheuren also points out the problem of the decline of public cooperation. However, when all the bodies are counted, what figure makes a successful census. In 1980, 98.6 percent of the population was counted. Let's say that this time 96.8 percent of the population is accounted for. Does this make the decennial census effort a failure? This will depend upon the differential undercount. Should we begin to find other ways to reach people based upon this? We will still have for the very most part usable small area data to work with. Most decisions will not change at all if the response rate does not decline drastically. Perhaps adjustment will adequately solve much of the undercount problem.

We should certainly accept the possibility of the need for "paradigm shifts". But there seems to be a problem. The paper tells us that the rolling sample approach and the use of administrative registers just won't do the job that's needed, and all things being equal, might even be more expensive.

If accurate data are needed not only for redistricting and reapportionment, but the allocation of funds for over 100 federal programs, and if local communities need information to update their plans and allocations, you immediately have to fall back on some intensive decennial census activity. And what about private sector uses? Correct market decisions based upon detailed information is still what pays the bills, including the tax bill. If you eliminate detailed information for local areas, efficiency will decline, which is something we as a nation cannot afford.

As Fritz knows, I'm a very strong supporter of the use of administrative records for making intercensal estimates. And it has been shown in this country, and in Canada as the next paper shows, that excellent work can be done in linking administrative data sets.

Therefore I believe that our best approach so far is not to throw out the present paradigm. Instead, we have to find ways to convince the American people that they have an important stake in knowing what their about. Further, we have to convince the policy makers that once in ten years is far too infrequent, a point that Scheuren makes quite well. Also, we mustn't abandon the concept of a quinquennial census, and we have to convince the policy makers that more intercensal work is needed.

For the first time in many years, you, the statistical agencies have a special opportunity. Over the years, there has been no one in very high circles who had a real interest in statistics and was also close to the decision makers. At present, the Chairman of the Council of Economic Advisors has the ear of the president. We know that he believes in the need for timely accurate data. Therefore, the Federal statistical system needs his support and you should ask for it.

On to the Leyes paper, which was a pleasure to read. This paper portrays a cogent attempt to build a file over time which will eventually yield excellent information between census efforts. However, the Canadian Privacy Act seems to limit the use of these data.

Statistically, however, this is kind of model where different data sets are linked, that we in the United States should explore to make better intercensal estimates. Perhaps this is where the paradigm shift should take place. Finally, I wonder if the private sector in Canada has taken advantage of these files for marketing purposes? How does the private sector in Canada interact with these data, if at all?

Two points on both papers. First, both discuss the inability to generate household information. I think that this would be harmful to both the public and private sectors, and I urge more work be done to solve this shortcoming.

Second, the private sector has developed many linked files, some good, some bad. There are claims that over 80 million households can be reached with at least one of these files, and demographic data are attached to these files. I suggest that your agencies, at the very least, learn what has been done in the private sector and maybe take advantage of it by getting us all together and sharing our knowledge.

Session 3
SURVEY COVERAGE EVALUATION

CONTROL MEASUREMENT, AND IMPROVEMENT OF SURVEY COVERAGE

Gary M. Shapiro¹
Bureau of the Census

Raymond R. Bosecker
National Agricultural Statistics Service

I. Introduction

Coverage errors can cause serious biases in estimates based upon sample survey data. Undercoverage may be substantial in many surveys, especially of selected subpopulations. For example, the estimated undercoverage of Hispanic males aged 14 and over is 23 percent in the Current Population Survey (Hainer et al., 1988). In economic surveys, new businesses may be missed at a higher rate than older ones. If the characteristics of the missed portion of the population are very different from those of the covered portion, serious biases in the survey estimates for the total population will result.

This paper is a condensation and editing of Survey Coverage, Statistical Policy Working Paper 17 (U.S. Office of Management and Budget, 1990). The 115-page working paper was prepared by the Subcommittee on Survey Coverage of the Federal Committee on Statistical Methodology. Subcommittee members are Cathryn S. Dipbo (Co-chair), Gary M. Shapiro (Co-chair), Raymond R. Bosecker, Vicki Huggins, Roy Kass, Gary L. Kusch, Melanie Martindale, and D.E.B. Potter. Robert Casady, Charles Cowan, John Paletta, and Richard Pratt also wrote parts of the working paper. This paper has numerous unattributed quotes from the full working paper. Although the authors of this short paper accept responsibility for all errors, credit for the good ideas and concept of the paper belongs to all subcommittee members. We would also like to thank Melanie Martindale and Vicki Huggins for their useful comments on this paper and Cora Wisniewski, Sue Chandler and Bessie C. Johnson for their typing.

The purpose of both this paper and the full report is to heighten the awareness of survey program planners and data users concerning the existence and effects of coverage error and to provide survey researchers with information and guidance on how to assess and improve coverage in sample surveys.

¹This paper is a condensation of Survey Coverage, Statistical Policy Working Paper 17. Authors are listed in the second paragraph. The views expressed are those of the authors and do not necessarily reflect those of their agencies.

This report utilizes a broad definition of coverage error. This is defined to include all possible sources of error which are not classified as observational or content errors (U.S. Department of Commerce 1978).

Section II of this paper discusses selected major sources of coverage error. IIA discusses errors which might occur before the first stage of sampling and IIB those that might occur after the first stage. Issues associated with the creation and maintenance of sampling frames, the choice of sampling frame and strategy, field listing and interviewing are included. Section III discusses selected methods for preventing, reducing and evaluating coverage errors.

II. Major Sources of Coverage Error

A. Sources of Coverage Error Before Sample Selection

(1) Conceptual Issues -- The importance of thinking carefully about the research goals, concepts, and targeted population(s) for a survey cannot be overemphasized. Coverage errors can be inadvertently designed into a survey from the beginning by incorrect specification of the concepts to be measured or the population(s) to be targeted by the survey. Vague definitions of populations and concepts tend to create coverage errors because they lead to inappropriate unit inclusions on, or exclusions from, a frame and even to naming a population which cannot be adequately represented by a frame.

(2) Frame Construction -- Once a decision is made concerning the target population, either the sample design must be based upon available sampling frames or a frame must be constructed specifically for the study. Dalenius (1985) notes the following three important properties of a frame:

- o Makes it possible to compute estimates concerning a population which is sufficiently "close" to the target population.
- o Serves to yield a sample of elements which can be unambiguously identified.
- o Makes it possible to determine how the units in the frame are associated with the elements of the (sampled) population.

The first stage of sampling is usually dependent upon a frame consisting of a physical listing of units. This may be a list of names of individuals, establishments, institutions, counties, cities, streets, etc., or a list of numbers attached to city blocks, land area segments, houses, pages, or any number of unique,

definable entities. However, as Kish (1965, p. 53) notes, a "Frame is a more general concept: it includes physical lists and also procedures that can account for all the sampling units without the physical efforts of actually listing them." Deming (1960) cites one exception to a list of units. This occurs when a watch is used to sample time intervals during which customers leaving a store are interviewed.

The units listed in the initial frame may not correspond to the units about or from which information is sought. Often, additional frames are needed for successive stages of sampling in order to progress from available sampling units to the units to be contacted or measured. For example, areas may be selected from a listing or array of all blocks in an area frame. Housing units inside sampled areas may then be listed and sampled in order to achieve a listing of persons to be sampled that are members of the target population from which information is sought.

A more complex example is the procedure for selecting items to be priced in the Consumer Price Index. The sample of priced items is selected from items sold by a sample of outlets which, in turn, was selected from a list of outlets created from information provided by interviews with consumer units in addresses sampled from the decennial census, new construction permits, and area listings. In this case, interviews are conducted in a sample of housing units to create a sample frame of establishments, not a population frame, from which a sample is selected. Within the sample outlets, probability methods are used to select increasingly more detailed classes of goods until a particular item is selected. A complete list of all the items available for sale is never constructed.

(3) Frame Errors -- Kish (1965) states that a "frame is perfect if every element appears on the list separately, once, only once, and nothing else appears on the list," and classifies possible frame errors into four types: missing elements, clusters of elements appearing on the list, blanks or foreign elements, and duplicate elements.

Missing elements is the frame error which causes greatest concern. Because they are missing, no examination of the sample from the frame will reveal the nature of that component of the population. Often, conclusions are erroneously extended beyond an incomplete frame on the tenuous assumption that missing units are like or very similar to those represented on the frame.

The initial sampling units may contain clusters of subunits which must be incorporated into the sampling design. An example is a list of farm operator names of which the vast majority represent a one-name/one-farm relationship but some represent a one-name/multiple-farm relationship. In this situation, there is a distinct

possibility for coverage error unless the interviewer has been thoroughly trained.

If a frame is created or an existing list modified for a particular one-time survey, elements on the list which are blank or are not members of the population of interest should be removed. If they are not removed, those appearing in the sample must be identified and properly handled in the survey process.

Duplication of units on the frame may result in overcoverage, i.e., some members of the population are represented more than once. Population totals may then be overstated and means could be biased.

4) Frame Maintenance -- Frame maintenance procedures are discussed as they relate to the classes of coverage error just described. These procedures can be classified as follows:

- o Adding new frame elements or births,
- o Eliminating or identifying inactive frame elements or deaths,
- o Correcting misclassified frame elements,
- o Identifying existing frame elements no longer in scope, or in scope for the first time, and
- o Determining whether or not elements have combined with other elements or have split from existing elements (e.g., change in ownership, mergers, and divestitures in an economic setting).

When the research population is dynamic, it is important that a frame which represents it be updated to reflect births. Section III discusses several methods for doing this.

The failure to identify deaths on a sampling frame does not necessarily imply a bias, since any deaths sampled would be representative of the universe of deaths. But, biased sample estimates can result if an inactive element is sampled and imputed for when no response is obtained.

A problem associated with many frames is not that elements are missing, but that they are misclassified or are not classified at all with respect to one or more variables. This assumes importance if the variable or variables that are misclassified determine either the elements eligible for sampling or the subpopulations for which estimates are produced. Housing occupancy status (vacant or occupied), geographic codes, SIC codes, etc., are examples of such variables.

Closely related to the problem of misclassification is the problem of out-of-scope elements, i.e., elements that if properly classified would not be part of the universe of interest. As with death elements, the presence of out-of-scope elements on a sampling frame does not result in any biased sample results should they be sampled (assuming the sample process identifies them as out-of-scope).

The composition of elements comprising a frame will often change over time. This is especially true for economic-based frames, where, for example, individual plants are bought and sold by companies, two or more companies merge, or companies divest. From a coverage point of view, ownership is important because the continued sample status of a sold establishment often depends upon the status of the buying company.

B. Sources of Coverage Error After Sample Selection

The full Survey Coverage report discusses three broad kinds of error occurring after the initial selection of a sample from a frame: (1) Incorrect association of sampling to reporting unit; (2) editing errors; and (3) other nonsampling errors. We discuss only the first of these in this paper.

Misclassification of occupied housing units as vacant units is a frequent type of classification error in household surveys. In many surveys, the population of interest consists of occupied housing units, but the frame consists of other types of units as well. In the Current Population Survey (CPS), for example, an interviewer is generally given specific addresses for interview. When an interviewer is repeatedly unable to find anyone home at an address (s)he must classify it either as a vacant noninterview (out of scope) or as a noninterview unit occupied by persons eligible for interview. In October 1966, the CPS reinterview concentrated on measuring this type of coverage error (U.S. Bureau of the Census 1968). This research revealed that more than 10 percent of the units classified as vacant were actually occupied by eligible persons.

In two separate evaluation projects in the 1970 Decennial Census, 11.4 percent and 16.5 percent of the units initially enumerated as vacant were misclassified (U.S. Bureau of the Census 1973).

We believe that error in listing persons within interviewed households (within-unit) is the most serious source of coverage error occurring after sample selection. Alexander (1986) has estimated that within-unit error results in overall undercoverage of four percent for persons 12 and over in the National Crime Survey. Within-unit error is probably more serious for blacks and Hispanics. Hainer et al (1988) point out that in the CPS, black

female undercoverage is close to the overall undercoverage of seven percent, but black male undercoverage is about 20 percent, suggesting that most of this undercoverage results from within-unit error.

There are several instances in which authors have speculated on large biases caused by within-unit error. One example of this is discussed by Hainer et al. (1988): "... Cook (1985) presents evidence suggesting that the National Crime Survey may underestimate the number of gun assaults by as much as one-third. He offers the explanation that the National Crime Survey does not adequately cover the kinds of people criminologists believe are most likely to be involved in the life of the streets (including participation in criminal activity...)" (Cook 1985, see also Martin 1981).

Hainer, et al. (1988) discuss at length the ethnographic research that has been done on household survey coverage. They suggest there are two main causes of respondent reporting error resulting in missed persons:

- o Some people, especially black and Hispanic males, are deliberately omitted because of potential loss of household income if their presence in the household were known to authorities.
- o There is a lack of correspondence between survey definitions of household residency and how people actually live.

III. Methods for Dealing with Coverage Errors

The previous discussion focused on sources of coverage error in selecting and maintaining sampling frames. Solutions to problems arising from the limitations of available frame sources are a major challenge to the survey design statistician. Some options, however, are available for dealing with coverage error. The options discussed are: Questions to specify concepts, current sampling frame, updated frame for births, random digit dialing, multiple frames, reinterview, estimation procedures, and evaluation methods.

A. Preventing Incorrect Specifications of Concepts

To avoid coverage errors caused by incorrect specifications of concepts, it is useful to ask a series of questions:

- o To what population(s) of units does this problem refer?

Distinguish among populations about which information is sought, those which will be frame units, and those which may be reporting units, if different from the frame units. For example, suppose one wished to do research on "the scholastic achievement (as measured by grades) of children of recent immigrants." In this case, "children of recent immigrants," more suitably specified perhaps as "persons aged roughly 5 to 17 enrolled in Grades 1 through 12 of the U.S. public schools and living in a household in which at least one related head has been resident in the United States 5 or fewer years," would be the population about which information is sought. However, it seems likely that one might need to construct two or more frames in order to reach this population. One of the frames might have U.S. public schools as units, while another might consist of residential addresses to be screened. In this example, reporting units might well consist of two groups, school recordkeepers and parents or guardians.

- o Is (are) this (these) population(s) observable or potentially measurable? How?

Continuing from the example above, one can see that the suggested specification of "children of recent immigrants" takes account of some of the presumably unobservable "children of recent immigrants", such as those who may be homeless and those who may not be currently enrolled in school. Among recent immigrants, those who entered the country illegally may not be observable, as well as those who died following entry, leaving school-age dependents. Sources for obtaining U.S. public schools and residential addresses might be lists from various agencies. Thinking through all possible categories of the populations of interest should reveal those subsets which cannot be measured or reached; those whose measurement (observation) might be achieved; and those which seem reachable with some existing or proposed methodology. Thus, the "children" may be reached by means of a household survey, school survey, and/or institutional survey (hospitals, orphanages).

- o Are there one or more subsets of this (these) population(s) which cannot be measured/observed in some way? What are these? Would they ever be measurable?

Continuing the example of "children of recent immigrants," some of the unobservable components of the populations discussed have already been mentioned. The potentially measurable components might be those who cannot be reached now but who might be reached using a methodology that is prohibitively expensive, such as scanning all death certificates or other sources of information to identify deceased recent immigrants. Thus, it may be useful to distinguish the inherently unobservable from the practically unobservable components of populations of interest.

- o Does time enter into the answer to one or more of the questions above, in the sense that the measurable population(s) may change or may have changed?

Continuing the example of "children of recent immigrants," one may find that a change in a legal boundary or definition can turn "internal migrants" to "recent immigrants" or vice versa. This would happen, for example, if Puerto Rico became a U.S. state, thus solving the problem of how technically to classify migrants to the mainland, who would become "internal migrants". Such a change might force a redefinition of the size and location of the populations of interest.

- o Have previous efforts been made to build a frame of this (these) population(s)? What problems were encountered in frame construction? Was one of these faulty conceptualization? Which of these problems has been solved?

This series of questions focuses on the need to locate previous research, to attempt to contact those who designed and conducted the research, or to obtain procedural histories about it and to evaluate carefully the definitions and language used by others. An assessment of previous research often reveals use of frames built for other purposes by still earlier researchers, especially when the frames are very expensive to assemble. Information needed for adequate frames may now be available (such as improved school lists) due either to improvements in information processing or to changes in laws regarding availability of administrative data.

B. Current Frames and Updating Old Frames

Use of old frames can result in serious coverage problems because births may be partially or totally excluded and other units may be misclassified. An obvious but important solution is to use current or recently built or updated frames whenever possible.

When an old frame must be used, it is important to have updating procedures to include births. One effective method for detecting new units is to periodically canvass the existing frame elements. As an example, all of the larger multiunit companies and some of the smaller companies on the Standard Statistical Establishment List are canvassed on a yearly basis. Companies are questioned as to whether or not they have started new operations.

A second method of identifying new units results from coverage maintenance operations performed for samples selected from the frame. As part of the questionnaire administration process in nearly all surveys, inquiries are made about the status of the sampled units and whether any changes in their status have occurred

since the last data collection period. Although the inquiries are targeted to sampled units believed not to be births, sometimes incidental information about other units (including births) can be obtained.

Several methods can be used for including new units in household surveys. The Bureau of the Census includes most new housing starts in its household surveys by sampling from building permit files. This is an efficient procedure, but building permit files do not identify illegal new construction, conversions, and new mobile home placements; nor do they identify new special places, such as dormitories, fraternity houses, boarding houses, and public housing. To illustrate, it was estimated for the 1985 American Housing Survey that approximately 25 percent of all new mobile homes were missed (Schwanz, 1988).

C. Random Digit Dialing

One household sampling method employed in an attempt to avoid omission problems is random-digit dialing (RDD) (Waksberg, 1978). The use of telephone directories as sampling frames often results in unacceptable levels of undercoverage because they omit unlisted numbers for some nontypical portions of the population. With RDD, a sample of telephone households is located through the use of randomly generated telephone numbers. In this way only those households without telephones are omitted. For many surveys, this could be considered a trivial exclusion. In others, differences between telephone and nontelephone households may have a profound impact on the characteristics being measured. For example, measures of poverty and income from entitlement programs would most likely be biased.

D. Multiple Frames

Coverage may be improved through the use of multiple frames. Sometimes no single frame fully covers the target population and merging independent source lists would be impractical. In this case separate probability samples from different frames can be used to expand coverage beyond any available single frame.

The application of overlapping multiple frame sampling most commonly found in Federal surveys is the use of an area frame and an overlapping list frame. The area frame is generally designed to provide complete coverage by including all U.S. land parcels as sampling units. The list frame is nearly always incomplete (a common attribute of lists), but its use provides certain sampling efficiencies which enable the multiple frame survey to provide the same precision at a much lower cost than would an area frame survey alone.

E. Reinterview

Reinterview can often be profitably used for both evaluation and control of coverage error. In the CPS, the regular reinterview program is able to detect misclassification of occupied housing units as vacant units, errors made in listing housing units in area segments, and errors made in missing persons within interviewed units. However, the CPS reinterview program serves many purposes and consequently fails to detect a number of these errors. A special intensive coverage check was done in the 1966-67 CPS reinterview. This check was much more successful than regular reinterview in detecting vacant unit misclassification and area segment listing errors, but still found few instances of within-unit errors (U.S. Bureau of the Census 1968).

A type of reinterview can also be used for nonresponse follow-up. A subset of original noninterviews can be more aggressively pursued to obtain complete or at least partial interviews, or alternatively, refusal households can be sent a very brief mail questionnaire asking why they refused and collecting basic demographic information.

F. Estimation Procedures

Estimation procedures may also be used to decrease the bias of survey estimates relative to the target population. One such procedure is the use of ratio estimation or benchmarking. The Bureau of Labor Statistics employs a benchmarking procedure to revise monthly employment estimates from the Current Employment Statistics survey. (U.S. Bureau of Labor Statistics 1989) Sample estimates are compared each year with later summarizations of mandatory UI reports filed by employers. The UI data, which serve as a benchmark, are an aggregation from the same source as the micro-data used to construct the frame from which the sample was selected, except that the benchmark data are one year newer. Hence, the benchmark file takes into account new firms or changes in industrial classification to ensure more accurate coverage. The completeness of the UI administrative data affords the opportunity to analyze and adjust for frame deficiencies (Thomas, 1986).

G. Macro and Micro Level Evaluation

Evaluation methods to independently determine the representativeness of the sampling frame(s) used are very useful for quality control. One method of measuring the degree of frame coverage error is comparative analysis. Comparative analysis can occur at two levels. The first is a macro level evaluation, which compares known population values with totals derived from summing characteristics for each sampling frame unit. The second type of analysis is performed at the micro or individual sampling unit

level. This most often involves matching of data available from different sources for individual units.

The Bureau of the Census utilizes a macro-level approach for frame completeness evaluation called demographic analysis. With this method, demographic data from various sources are used to develop expected values for the population as a whole and by race, age, and sex to compare with the census counts.

On a micro-level basis the Bureau of the Census matches census returns against administrative records for drivers' licenses from State departments of motor vehicles and against registers of resident aliens supplied by the Immigration and Naturalization Service.

IV. Conclusion

This paper has presented many of the major points treated in the full Survey Coverage report, whose purpose is to provide information about the types and effects of coverage error in surveys and guidance on how to assess and improve survey coverage. We found few studies, however, which actually measure coverage errors in surveys and even fewer which address the impact of coverage error on survey estimation. The paper implies that significant resources should be allocated to the conceptual and planning stages of surveys, and that procedures providing for the evaluation of coverage and for minimizing and controlling coverage error be clearly established and included in the survey design.

As to the seriousness of coverage error, the largest single source of coverage error identified in the full Survey Coverage report for an economic survey is a 20 percent underestimate in the 1988 Economic Census statistic of receipts for nonemployer establishments due to misclassification. For household surveys, large single source of overall coverage error is an estimated 4 percent undercoverage in the National Crime Survey estimates of persons aged 12 and over due to within housing unit listing errors. (Undercoverage from this source for some subgroups is much worse.) Since we know that single sources themselves can be significant, the overall effect of all sources of coverage error on survey products is of great concern.

Several leading methods for identifying and assessing coverage error and for improving coverage have been mentioned here. The full report treats these and other methods in detail. It also provides case studies of specific Federal surveys which illustrate various frame and coverage issues.

The methods that apply to most surveys and which can lead to significant improvements in data quality are the use of multiple

frames to improve coverage at the sampling stage and weighting adjustments to reduce bias from coverage error.

References

Alexander, C. (1986), "The Present Consumer Expenditure Survey's Weighting Methods," in Population Controls and Weighting Sample Visits, Washington, DC: U.S. Bureau of Labor Statistics.

Cook, P. (1985), "The Case of the Missing Victims: Gunshot Woundings in the National Crime Survey," Journal of Quantitative Criminology, 1, 91-102.

Dalenius, T. (1985), "Elements of Survey Sampling," Notes prepared for the Swedish Agency for Research Cooperation with Developing Countries (SAREC).

Deming, W. (1960), Sample Design in Business Research, New York: John Wiley and Sons, Inc.

Hainer, P., Hines, C., Martin, E., and Shapiro, G. (1988), "Research on Improving Coverage in Household Surveys," Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census, pp. 513-539.

Kish, L. (1965), Survey Sampling, New York: John Wiley and Sons, Inc.

Martin, E. (1981), "A Twist on the Heisenberg Principle: Or, How Crime Affects Its Measurement," Social Indicators Research, 9, 197-223.

Schwanz, D. (1988), "1985 Type-A Unable-to-Locate Rates for the AHS National Unit Samples," Internal memorandum, U.S. Bureau of the Census.

Thomas, A. (1986), "BLS Establishment Estimates Revised to March 1985 Benchmarks," Washington, DC: U. S. Bureau of Labor Statistics.

U.S. Bureau of the Census (1968), "The Current Population Survey Reinterview U.S. Program, January 1961 through December 1966," Technical Paper 19, Washington, DC: U.S. Government Printing Office.

U.S. Bureau of the Census (1973), "The Coverage of Housing in the 1970 Census," Report PHC(E)-5, Washington, DC: U. S. Government Printing Office.

U.S. Bureau of Labor Statistics (1989), Employment and Earnings, 36 (12).

U.S. Department of Commerce (1978), "Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics," Statistical Policy Working Paper 4, Washington, DC: U.S. Government Printing Office.

U.S. Office of Management and Budget (1990), "Survey Coverage", Statistical Policy Working Paper 17, Washington, D.C.

Waksberg, J. (1978), "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, 73, 40-46.

QUALITY OF SURVEY FRAMES

Judith T. Lessler
Research Triangle Institute

1. Introduction

This paper focuses on the quality of sampling frames with particular emphasis on the relationship of the sampling frame to the overall error of survey estimates. It also presents some examples from studies that have been conducted by the Research Triangle Institute (RTI).

The frame is a fundamental element of scientific survey research. Probability sampling involves selecting a subset of units from a finite collection of units in a manner that lets one determine the probability of obtaining that subset. The sampling frame is the finite population of units to which the probability sampling mechanism is applied. Thus, the type of frame used for a survey and any deficiencies or inefficiencies in it affect the total error of the survey estimates.

2. Definition of a Frame

The population of frame units is not necessarily equivalent to the population for which information is to be collected. In this paper, I refer to the population the survey researcher wishes to make measurements on as the target population and the individual components of that population as elements. This population may not be the same as the inferential population. For example, the National Human Monitoring Program of the U.S. Environmental Protection Agency (EPA) conducted a special study of mirex residues in human adipose tissues (Leininger et al., 1980). Mirex is a persistent insecticide that has been used to control fire ants. Human adipose tissue specimens were collected from selected surgical patients and cadavers and chemically analyzed for the presence of mirex residues. The inferential population in this study was not the sick and the dead, but, rather all persons living in the areas subject to application of the insecticide.

Just as the target population is not necessarily the same as the inferential population, neither is the population of frame units the same as the population of target elements. Thus, noting this distinction and the role of the frame in survey sampling, I once defined a frame as follows:

"The frame consists of materials, procedures, and devices which identify, distinguish, and allow access to the elements of the target population. The frame is composed of a finite set of units to which the probability

sampling scheme is applied. Rules or mechanisms for linking the frame units to the target population elements are an integral part of the frame. The frame also includes auxiliary information (measures of size, demographic information) used for (1) special sampling techniques such as stratification and probability proportional to size sample selections, or (2) special estimation techniques, such as ratio or regression estimation."

I like this definition because it clearly recognizes that different types of frames support different types of sampling and estimation procedures.

However, I think that it fails to recognize a key aspect of sampling frames, namely, the types of measurement designs they support. To illustrate, if a survey is to be conducted by asking questions or by gathering information from records, the reporting units are not always equivalent to the target elements. For example, suppose we wanted to know the family income of all children who attended the Saturday afternoon swimming classes at Sometown Community Park. A sampling frame consisting of a list of all swimming classes and the times that they met would provide us easy and efficient access to the target population of children. We could go to the class and identify each child; however, this would not be very helpful because few children know their family incomes. Thus, we need to insert a key word in the above definition -- measurement -- yielding:

"The frame consists of materials, procedures, and devices which identify, distinguish, and allow access to and measurements on the elements of the target population. The frame is composed of a finite set of units to which the probability sampling scheme is applied. Rules or mechanisms for linking the frame units to the target population elements are an integral part of the frame. The frame also includes auxiliary information (measures of size, demographic information) used for (1) special sampling techniques such as stratification and probability proportional to size sample selections, or (2) special estimation techniques, such as ratio or regression estimation."

3. Components of Quality

Researchers who are choosing a sampling frame for a survey need to consider a number of factors when making that choice. These include:

- coverage of the target population

- efficiency of the sample designs that are supported by the frame
- effect of the frame on nonresponse errors
- types, costs, and quality of the measurement designs supported by the frame
- cost of constructing the frame
- accuracy of information on the frame

Coverage of the target population: It is widely recognized that several aspects of a sampling frame can cause bias in survey estimates. Missing target elements, inclusion of nontarget population elements, unrecognized multiplicities, and failure to account for the clustered nature of frames during sampling and estimation can all introduce bias in survey estimates.

Efficiency of sampling and estimation: The structure of the frame, the information it contains, and the quality of that information will determine the types of sample designs and estimation procedures that can be used in a survey. Simple frames lacking auxiliary information support simple sample designs; complex frames containing auxiliary information support more complex designs, which are generally more efficient. Frames used for sampling business establishments are a good example. Lists that also include information on the size of the establishment will permit sample designs that are much more efficient than those that could be designed using a simple listing of establishments.

Effect of the frame on nonresponse errors: The type of frame that is chosen also has a major impact on nonresponse errors. Often, a frame that provides efficient access to large segments of a target population will also be guarded by "gatekeepers" who can deny access to the target elements. For example, if one would like to conduct a survey of young people aged 12-17, using a school based sampling frame rather than an area household frame will provide more efficient access to the great majority of this target population. To use such a frame, one usually needs permission from school district personnel who can, in a single decision, deny access to large segments of the target population. In a national survey, failure to obtain cooperation from the large city school districts can have a devastating impact on our ability to control nonresponse errors.

Types of measurement designs supported/cost of making measurements: The frame that is chosen for the survey also affects the types of measurements that can be made. Frames of telephone numbers using random digit dialing provide access to a very large part of the household population. Using this frame, however generally limits one to making measurements by asking questions.

One cannot weigh the person or collect a blood sample although one could, of course, obtain the person's address and make the direct measurements in subsequent visits. These subsequent visits would cost more than using an area housing unit frame because the sampled elements would be widely dispersed.

RTI recently completed a survey for the Food and Nutrition Service of participants in WIC (Women, Infants, and Children Feeding Program). Much of the information that needed to be collected could be abstracted from the WIC records; however, other information required an interview with the WIC participant. A sampling frame that consisted of lists of WIC agencies and lists of persons served would have been the most efficient for collecting the record data; however, it would have been very inefficient for conducting the interviews. Because of this, we developed procedures for listing people as they arrived at WIC clinics for their initial enrollment into the program.

Cost of constructing the frame: When assessing the relative quality of various sampling frames, we must consider the cost of constructing the frame. A frame that includes "size measures" for the units may be permit more efficient sampling; however, it may be too costly to determine the size of the units. The money spent on constructing the frame might be better spent in increasing the sample size.

Accuracy of information on the frame: If the auxiliary information on the sampling frame is inaccurate, the efficiency of sample designs and estimation procedures that make use of this information will be reduced.

4. Examples

RTI has conducted many types of surveys using many kinds of sampling frames including area household surveys and random digit dialing surveys, as well as surveys of schools, businesses, military personnel and families, nursing homes, hospitals, and so on. We also do a number of environmental surveys, and I will describe two of these to illustrate the points discussed earlier.

4.1. Of Flowing Waters

The first example shows how a frame can influence in several ways the quality of a survey's estimates. The goal of the 1982 National Fisheries Survey was to measure the biological quality of the Nation's flowing waters. After some discussion of exactly what was intended by the phrase, the "Nation's flowing waters," the statisticians on the project turned to the task of developing an operational definition of sampling units and target elements for use in the survey. It turns out that the EPA has developed a cataloging system in which each body of water in the United States

is segmented into well-defined units called reaches, described according to the following definition (Horn, 1981):

"Most reaches represent the approximate centerlines of streams and extend between points of confluence with other streams. The reaches constructed within open waters are generally straight lines connecting tributary streams with assumed transport paths through the open waters."

In addition, the U.S. Geological Survey (USGS) has a system in which the United States is divided into nonoverlapping areas based upon the configuration and sizes of watersheds. There are 2,100 cataloging units (CUs) contained in larger regions called water basins or hydrologic regions. When we designed the survey, EPA maintained a River Reach File that contained some 68,000 reaches defined within these CUs. This file was not complete because it was estimated that the total number of reaches was around 179,000. Moreover, a clustered design was not needed to control data collection costs because the survey was to be conducted by mailing questionnaires to local fisheries biologists who were familiar with each waterbody. In addition, a very accurate (but costly) digitizing procedure for identifying reaches and for measuring their length was available. Thus, staff decided to select the sample in two stages: (1) sampling CUs, then (2) reaches within CUs using maps to identify the reaches.

We established the following operational definition of the target population:

All reaches of rivers and streams that were:

- a. contained in the 48 contiguous States;
- b. shown on 1:500,000 USGS maps;
- c. including watercourses shown on the maps as being seasonally intermittent, impoundments, reservoirs, canals and constructed channels, and waterways; and
- d. excluding the Great Lakes and other lakes, marine waters, estuaries, and wetlands (Glauz, 1984).

One interesting feature of this definition is the specification of the map scale. The scale 1:500,000 is in inches -- one map inch for every 500,000 inches. Because reaches are defined by points of confluence, maps with higher resolution would show more reaches and maps with lower resolution fewer reaches. Smaller-scale maps were not available; thus, our definition of the target population was limited by the materials we had available for identifying its elements (given the available budget).

Measures of size were constructed for the first-stage sample by obtaining maps of all the 2,100 CUs and measuring the length of all the eligible waterways using a map meter. Grids were drawn on the maps to facilitate keeping track of the measurements, and cataloging units were randomly assigned to the staff performing the measurements.

A first-stage sample of 302 CUs was selected with probabilities proportional to size. Within this first-stage sample, a second-stage frame was constructed using automated digitizing equipment to trace, list, and record the size of each reach in the 302 selected cataloging units. A total of 1,303 reaches were selected from this second-stage frame.

This example illustrates several ways in which the frame influenced the quality of the survey estimates. First, the materials and procedures that could be afforded for constructing the frame limited the target population to reaches that were visible on the 1:500,000 scale maps. Smaller reaches could have been identified by selecting a sample of areas and using a counting and listing procedure; however, the budget for the survey did not permit such an activity. Second, the use of size measures for selecting first- and second-stage sampling units increased the efficiency of sampling. Third, the cost of constructing a complete list of all reaches required the use of a two-stage design.

4.2. Of Passing Time

The second example illustrates the relationship between the frame, the definition of a target population, and the measurement design. RTI recently completed the National Alachlor Well Water Survey (NAWWS) that required distributing a sample in both time and space (Whitmore et al., 1990). The goal of the survey was to estimate the frequency of occurrence of the herbicide alachlor in private rural wells used for domestic consumption. Because the water in wells is not static, sample wells could not be monitored at arbitrary points in time without introducing an unknown temporal bias into the sample. Data were to be collected over a 1-year period; thus, one of the first tasks was to decide on a definition of the target population by dividing the year into units for which it was possible to collect measurement.

A major constraint on the choice of a time period for the survey was the amount of time it would take to make a measurement. A year into months, weeks, days, hours, minutes, and so on. The lower limit would be the time required to draw and package the amount of water required for an accurate chemical analysis from the well -- a few hours. Partitioning the year into hours and selecting a sample of hours, however would have required a survey team to be at the well head standing at the ready while they waited

for the sample hour. In truth, the entire process of collecting water samples for the survey was much more complicated.

The survey team needed to contact the owner of the well, obtain his or her consent to draw water from the well, make an appointment with the resident (not necessarily the owner) for obtaining the water sample, travel to the site, identify a tap or hole for collecting the water (collection of water before any treatments was preferred), measure water temperature by running water through a flow-through cell, continue to run the water until a stable temperature was achieved or 10 minutes had passed, fill three large sample bottles, collect an additional water sample and mix it with a stabilizing reagent, and package the water bottles for shipping. In addition, observations and photograph(s) of the well site and surrounding area were needed as were questionnaire data on water use, well characteristics, and the surrounding area. After considering the time required for the survey teams to implement the entire measurement process, we decided that (with the resources available) dividing the year into observational units smaller than a month would not be feasible. Therefore, the target population for the survey was defined as well-months.

An assumption that underlay all NAWWS estimates was that the herbicide concentrations would be stable for the entire month. Dividing the year period into smaller units would have reduced measurement error; however, this would have also resulted in more missing data because the data collection team would have had severe difficulty in obtaining the measurements at the prescribed time.

To increase the chance that the concentrations were stable for the sample month, temporal strata were formed based upon ground-water recharge conditions. Prior information was used to classify each month into a historically low, medium, or high recharge stratum. Because the first-stage spatial sampling units were counties, temporal strata were created for each county.

References

Glauz, W.D. (1984) 1982 National Fisheries Survey. Volume II: Survey Design. FWS/OBS-84/14. U.S. Fish and Wildlife Service, Washington DC.

Horn, C. Robert (1981) The reach file: a digital base of streams and lakes (memo).

Leininger, Carol, Donna L. Watts, Charles Sparacino, and Stephen Williams (1980) Mirex Residue Levels in Human Adipose Tissue: A Statistical Evaluation. RTI Project Report, Contract 68-01-5848. U.S. Environmental Protection Agency, Washington, DC.

Whitmore, Roy W., et al. (1990) National Alachlor Well Water Survey. Volume I: Survey Design and Data Collection Final Report. RTI Project No. RTI/3895/04-03F. U.S. Environmental Protection Agency, Washington, DC.

DISCUSSION

Fritz Scheuren
Internal Revenue Service

Judith Lessler and Gary Shapiro and Ron Bosecker deserve our thanks today for their thorough "coverage of coverage." They have very ably reminded us of the important quality features of this aspect of a survey.

General

Taken together, the two papers provide a valuable summary of current practice. The papers complement each other nicely. In particular, we have been given two viewpoints today -- one, from the public sector and, the other, from the private sector of survey research. Differences in emphasis arise due to these perspectives. One example would be the degree to which frame construction is ad hoc (private sector) versus ongoing (public sector). More specifically, maintenance of frames is covered in detail in the Shapiro-Bosecker paper, but only touched on in the Lessler one.

A key issue in frame construction arises when we have a target finite population, but our real purpose is in making inferences about an ill-defined superpopulation. Judy's phrase "of flowing waters" says it all. Frame construction is part of learning what is already known before conducting a survey. It is part of connecting the measurement process with the "thing" to be measured. Coverage adjustments have this flavor of connection, too.

The cognitive research movement needs to be at least mentioned in the context of survey coverage issues, if only because of the conceptual challenges in defining the target population and the even more difficult challenge of "defining" the population of inference. Just look at the problem of within-household undercoverage, for example. Maybe Judy Lessler or our Chair, Cathy Dippo, would like to comment on these cognitive aspects, since they have been heavily involved in this emerging area.

Both speakers have constructed somewhat different taxonomies of survey coverage errors. One could profitably relate and refine their approaches; however, I found both useful as is.

On the whole, the papers do an excellent job of describing (albeit in broad terms) the main technological aspects of frame construction, maintenance and coverage. I have only one quibble: I was surprised by the complete omission of any mention of record linkage.

Finally, one last point of a general nature: the Shapiro-Bosecker paper should whet your appetite for the larger effort conducted by the Federal Committee on Statistical Methodology (FCSM). The FCSM subgroup led by Gary and Cathy Dipppo conducted an excellent series of case studies (Subcommittee on Survey Coverage 1990). These studies are, however, largely descriptive, rather than proscriptive -- a point I will turn to at the end of these brief remarks.

Quality

This two-day workshop is supposed to be about quality, so I would like to connect the present papers somewhat more to that theme than has been done already. In doing this, I want to shift the focus from PRODUCT quality to PROCESS quality and look more at how to improve the processes that we use to construct frames and conduct surveys.

At IRS, we are following an action-oriented quality management approach advocated by Juran (1986), Deming (1986) and others. This is in contrast to the mainstream statistical emphasis which has long focussed more on measurement and perhaps not enough on improvement. Anyway, Juran divides quality, like Gaul, into three parts:

- o Planning. -- The steps to be taken to prepare, including establishing the desired level of quality (implicitly or explicitly).
- o Control. -- The steps needed to implement and to achieve the desired level of quality.
- o Improvement. -- The efforts undertaken to make further improvements in quality over those initially planned.

Figure A provides a generic example giving you some typical steps taken at each of these three stages of quality management. This is an approach that we, at IRS, have begun to use to help the Census Bureau avoid a repetition of the 20 percent underestimate (for 1987) in the economic census statistic on receipts for nonemployer establishments -- among the largest coverage problems mentioned in the Shapiro-Bosecker paper (Greenia 1990; Konschnik and Moore 1990).

Conclusion

Let me conclude by making some recommendations on possible next steps for a follow-up to the fine FCSM efforts to study survey coverage quality issues:

- o Complete the learning from each of the FCSM case studies by subjecting them to a checklist like that in Figure A, to summarize for each case what the quality management steps were for survey coverage.
- o Choose the "best of the best" approaches. The Japanese word here is DANTOTSU. This (partly subjective) step is the beginning of an initial conjecture on a prescription for potentially system-wide improvements.
- o Use some of the results of this proscriptive exercise to initiate improvements and to gain (back) a deeper knowledge of the once-American-now-partly-Japanese ideas that surround the second quality revolution.

In the last session, I talked about paradigm shifts in census-taking. I am unable to resist doing so again. In particular, I would like to refer you to an excellent article in Scientific American (Gomory 1990) on two improvement paradigms: ladders and cycles. My belief is that a big -- or ladder -- paradigm shift (like cognitive methods) may not be needed in the coverage area (unlike in census-taking). But, whether it is or not, we must make better use of small -- or cycle -- paradigm shifts and learn faster from each other's successes (and failures). The Federal Committee's work, as summarized today by Gary and Ron, plus Judy's ideas, offers a platform for at least some of the improvements needed.

References

- Deming, W. Edwards (1986). Out of the Crisis, Cambridge: Massachusetts Institute of Technology.
- Gomory, Ralph (1990). "Of Ladders, Cycles and Economic Growth," Scientific American, June, 140.
- Greenia, Nick (1990). "Sole Proprietor/ship IMF/BMF Connection: An Application of the Juran Trilogy," Statistics of Income working paper (unpublished), Internal Revenue Service.
- Juran, J. M. (1986). "The Quality Trilogy," Quality Progress, American Society for Quality Control, Inc., August, 19-24.
- Konschnik, Carl A. and Moore, Richard A. (1990). "EC-14, A Study of the Methodology for Removing Employer Duplicates from the Nonemployer Universe for the 1987 Censuses of Retail and Services," Business Division internal memorandum (September 19), Bureau of the Census.

Subcommittee on Survey Coverage, Federal Committee on Statistical Methodology (1990). Survey Coverage, Statistical Working Paper 17, Office of Management and Budget.

Figure A.—THE QUALITY MANAGEMENT PROCESS*

JURAN'S BASIC QUALITY PROCESS	A GENERIC APPLICATION	
	EXAMPLE	COMMENTS
QUALITY IMPROVEMENT		
1 Organize for quality improvement.	Establish Quality Council, obtain management commitment, provide training resources	
2 Identify problems.	Unacceptable response, accuracy rates; late surveys; cost overruns; missing target elements, etc.	
3 Select problem.	Select doable, high-profile, high return on investment problem	
4 Analyze root cause (may require specialized statistical skills such as problem-solving process, design of experiments)	Collect, plot and graph data; perform statistical analysis	
5 Identify possible solutions and select solution.	Solution should eliminate root cause problem	
6 Test and implement solution.	Perform trial run	If problems (errors) are no longer a problem in a process, improvement efforts can be directed towards greater efficiency, new products or services, quality of work life, etc.
7 Track effectiveness (may require specialized statistical skills such as sampling, control charts and analysis.)	Exercise "Quality Control"	

Note The terminology used in steps 2 through 7 refers to a team approach to the usual scientific problem-solving process developed by Juran.

* Prepared by Otto Schwartz and Raymond Shadid, IRS Statistics of Income Division, R.S.P.
(Based on J.M. Juran, "Quality Trilogy," Quality Progress, August, 1986.)

Figure A.—THE QUALITY MANAGEMENT PROCESS*

JURAN'S BASIC QUALITY PROCESS	A GENERIC APPLICATION	
	EXAMPLE	COMMENTS

QUALITY PLANNING

1. Identify customers (internal or external.)	Government agency	
2. Determine customer needs.	Information: measurement	
3. Design products or services to respond to needs	Survey design	
4. Establish quality goals and minimize costs.	Accuracy; timeliness; cost	
5. Develop a process that meets needs.	Define population; construct sampling frame; determine sample size; develop data gathering process	The process involves people, methods, material, machines, environment and organization
6. Prove (test) process capability (may require specialized statistical skills.)	Carry out trial survey; simulate process	

Note: The process as designed generally does not fully satisfy customer needs. This may be due to planning deficiencies or to perceived resource constraints.

QUALITY CONTROL

1. Choose what to control.	Accuracy; timeliness; cost	Quality control must be exercised through continuous internal (self-policing) and periodic external (independent audit) means.
2. Choose unit of measurement.	Rate of response; accuracy of response; processing times	
3. Establish measurement.	Set up measurement process	
4. Establish standards of performance	Response rate; accuracy rate; time frames	Any "acceptable error rate" other than zero should be considered an interim goal.
5. Measure actual performance (may require specialized statistical skills such as sampling)	Perform measurement; prepare control charts, graphs, etc.	
6. Interpret difference	Compare actual performance to expected performance	
7. Take action on difference	Bring out-of-control conditions back into control	

Note: The process is expected to produce the "planned" quality level. If a problem occurs (a "sporadic spike"), reasons are determined and the process is brought back into control.

DISCUSSION

Joseph Waksberg
Westat, Inc.

1. Content of the Two Papers Presented

The two papers present a good review of issues relating to sampling frames. Their emphasis is on coverage but they are not exclusively devoted to coverage. They would be useful reading for anyone developing a design for a new survey, or reconsidering sampling and related methods for a continuing survey. Although much of the material in the two papers covers the same subjects, there is considerable difference in focus. As a result, the authors provide a well-balanced discussion of options normally available and considerations that should be kept in mind in choosing among alternatives. Shapiro and Bosecker mostly describe properties of frames that affect sample designs. Judy Lessler places more emphasis on how the frames can affect measurement methods, and conversely the way measurements can influence the choice of frames. The two papers thus complement each other nicely. The papers contain definitions, properties of frames, important problems inherent in some frames, and in some cases suggestions and recommendations for dealing with the problems. I'd like to discuss in more detail several of the points made in the papers.

2. Minimizing Total Means Square Errors

The authors of both papers imply, although they do not specifically say, that efforts to improve coverage by choice of suitable frame and procedures for working with that frame, are all part of attempts to minimize the total mean square error of survey estimates. Although the minimization usually cannot be done in precise mathematical form, it is almost always part of the background thinking in developing survey procedures. Judy Lessler discusses the relationship of the frame to measurement methods. In practice, the situation is even more complex, with frame, measurement methods, sample design, and sometimes estimation methods intertwined. All four frequently have to be taken into account in decisions on choice of frame and intensity of efforts to improve coverage. Let me give some examples:

- a. About 25 years ago, the sample design for the CPS and the other Census-conducted national population and housing surveys changed from using area sample frames to list samples in most of the U.S. The list samples consist of the set of addresses in the preceding census plus building permits issued for new construction since the census date. In considering pros and cons of the two types of frames, it was clear there were biases in both

systems. Building permits do not quite cover all additions to the residential stock of housing, even in areas requiring permits for construction. In addition, there is some loss because permits cannot always be located in the building permit office. Finally, in theory, the building permit frame should consist of permits for units constructed after the date of the Census. The time period is somewhat fuzzy and permits issued in the year or so preceding the census cannot be unambiguously classified on whether they were included in the census (at least not without an inordinate effort and cost). Area frames have other types of bias. The maps Census has used over the years are frequently outdated and many are difficult for interviewers to use. In addition, experience over the years indicated that interviewers cannot locate all units in area segments and a small loss consistently appeared. This undoubtedly affects the quality of the frame although how much and in what direction are difficult to quantify. However, one aspect of the comparison of two frames is quite clear. The list sample had a smaller variance. This is because over the 10 to 15 years following each census, the measures of size of the area segments became seriously out of date. Starting a few years after each census, the area segments became quite variable in size, and this variability increases progressively over the years. The list sample provides relatively consistent segment sizes. The change from area to list sample was mainly introduced to reduce the variance arising from variability in segment size. It appeared probable that coverage would also improve, although the evidence on this was weak.

- b. Westat has carried out three cycles of the National Survey of Family Growth for Health Statistics. The sample designs for first two were based on traditional area samples. For the third cycle, the National Health Interview Surveys (NHIS) was treated as the sampling frame, and the sample consisted of a subsample of eligible persons in the NHIS in the preceding year and a half. The original purpose of this revision in the frame was to reduce the cost of the extensive screening necessary to locate the required number of eligible persons. In order to keep the screening costs in the two earlier cycles in check, a complex sample design with variable sampling rates was necessary. The NHIS permitted the elimination of most of the variable rates resulting in substantial reductions in variances for many statistics. Although it was recognized that there would be a small loss in coverage from inability to locate some of the persons who moved after the NHIS interview, it was felt that the reduction in variances compensated for it. There was a side benefit to the procedure adopted. The

NHIS contained considerable data on social and health characteristics of the persons in the frame. This information was very useful in the nonresponse adjustment procedure.

- c. Random digit dialing (RDD) is, of course, much cheaper than face-to-face interviewing, especially when screening for a target population is necessary. The difference in cost is so great that except for the major complex national surveys requiring an extraordinary degree of accuracy and surveys requiring physical measurements, most surveys both in the government and private sectors are now carried out over the telephone. Although RDD is presumably only a sampling device and the sample persons can be interviewed over the telephone or in home visits, telephone interviewing is so much cheaper that researchers generally pick it. The frame thus influences the choice of measurement methods. It's interesting that the emergence of RDD has spurred research into the quality of telephone and face-to-face interviews, and the findings have made telephone interviewing a more respectable measurement method.

3. Narrowing Definition of Target Population

Shapiro and Bosecker mention that in some circumstances it is useful to narrow the definition of the target population to one that permits use of a more accessible frame. In some sense, this is almost always done. Surveys using area samples implicitly define the target population as those persons who are normally reported in area samples, thus excluding the undercoverage normally found. Business surveys frequently use businesses with one or more employees instead of all businesses, etc. I'd like to discuss two aspects of a narrower definition.

3.1. Risks of Narrowing Definition

I think most researchers would agree that the redefined target population should satisfy two criteria:

- a. It accounts for a very high proportion of the true target, preferably 85 to 90 percent or more.
- b. Characteristics relating to the subject of the study should not be wildly different in the narrower population and the missing piece.

The second criterion is quite important. It's not always recognized that even if the missing part is a small part of the

inferential population, in some cases it can have big effects. Let me give some examples.

RDD telephone surveys are probably the most common method by which a population is restricted to permit use of less expensive sampling and interviewing methods. About 93 percent of the U.S. population live in telephone households, so that the first criterion is satisfied. The extent to which the second criterion is satisfied depends on the statistic being studied. For example, in examining the feasibility of using RDD for a study of school drop-outs, the following results emerged¹. Figure 1 shows drop-out rates in telephone and nontelephone households for 14-21 year old youths. The shaded and cross-hatched boxes represent all drop-outs, and youths who dropped out in the past year. It can be seen that drop-out rates in nontelephone households are about five times the rates in telephone households. The discrepancy is large enough to substantially affect the total, even though the nontelephone households only account for seven percent of all household. In fact, estimates of drop-out rates from telephone households alone would understate the actual drop-out rates by about 25 percent. These estimates can be improved somewhat by post-stratifying the telephone household results, but they still seriously underestimate the true drop-out rates. Figure 2 shows drop-out rates for telephone households as a percentage of drop-out rates for the total population, and similar ratios when post-stratification is used to compensate for known deficiencies in using telephone households as a surrogate for all households. The post-stratification cells comprised single years of age, race/ethnicity, and highest grade attended by the head of the household. As can be seen, post-stratification improves these rates considerably. The ratio for total drop-outs goes from 77 to 85 percent, but the rates are still much below the actual numbers.

Telephone households showed up much better for other statistics studied in the same feasibility study. An analysis of enrollment in education programs for three- to five-year olds showed only trivial bias in restricting a study to telephone households. Figure 3 shows ratios of enrollment rates in telephone households to all households., As can be seen, post-stratification practically eliminates whatever bias exists in the data.

Thornberry and Massey² similarly report wide differences among health-related items in the extent to which telephone households can be considered to represent all households. For the vast majority of items, there is no problem, but problems exist for items related to income. For example, estimates of the number of persons with private health insurance would be overstated about four percent if it were based only on telephone households. Most other health items would be affected only slightly.

3.2. Population for Which Estimates Are Prepared

When a survey uses a frame that does not include the total target population, there should be a clear and unambiguous statement on how the sample was selected. However, the estimation methods should attempt to adjust the narrow population so that inferences can be made about the broader population. Some researchers feel that there is something wrong in expanding the results beyond the boundaries of the frame. I don't think it makes any sense to tell data users who are interested in a specific population, that because it's cheaper or easier you've done a study of another group and they can't infer anything about the population they're interested in from the study.

Of course, no one would make such a strong statement. However, there is an implication that the results tell you about the inferential population but as a scientist you're not allowed to say so. It seems to me that since the only reason for having done the survey was to shed light on the inferential population, it makes sense to do whatever is necessary to produce the best estimates you can for that population. This is, in fact, a commonly accepted procedure. The weighting or imputing procedures used to reduce nonresponse biases implicitly assume that one wants to produce statistics for the total rather than the respondent population. Similarly results of telephone surveys are usually inflated up to the level of the full population.

There are some real dangers in not taking the trouble to produce estimates for the inferential population. Let me cite an example where even the producers of the statistics forgot the statistics referred to a narrow population.

In November 1989, the Census Bureau issued a report on the Black population in the U.S. One of the statistics cited in the report was that the black female to male ratio was 100 to 88 compared to 100 to 96 for whites. The difference is startling, and if true has serious social implications. However, the text statement of this statistics is followed by a sentence which mentions that the ratios may be affected by greater census undercoverage of males than females. Elsewhere in the report is a footnote stating that the numbers reflect only the civilian noninstitutional population. The term "may be affected" is a gross understatement of the effect. If one takes coverage and institutional population into account, the discrepancy in the sex ratios between blacks and whites is cut by more than half. The full report gave no hint that the sex ratios are affected that much by these two factors. Furthermore, by the time a press release was issued by the Bureau of the Census, the fine line between the population actually covered in the CPS and the total population was lost, and the numbers were described as reflecting the difference between the total black and white population. The only way one can avoid these kinds of misinterpretations is to make the best

adjustments one can to have the data reflect the population that readers of the report assume is referred to.

References

1. J.M. Brick, and J. Burke, "Undercoverage Bias in the Field Test for the National Household Education Survey," report by Westat Inc. to the National Center for Education Statistics, 1990.
2. O. T. Thornberry and J.M. Massey, "Trends in U.S. Telephone Coverage Across Time and Subgroups," Telephone Survey Methodology, edited by R.M. Groves et al, John Wiley & Sons, 1989.

OMB-COPAFS Seminar on Quality of Federal Data

5

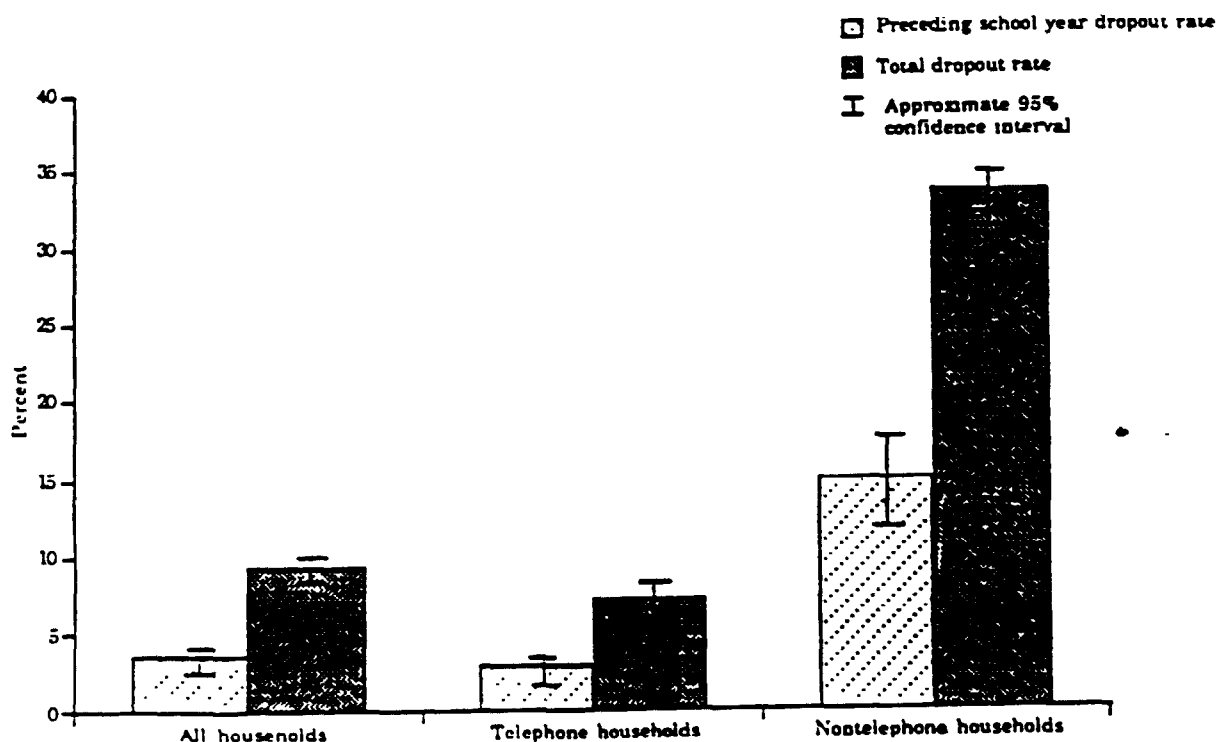


Figure 1. Estimated dropout rates by telephone status for 14- to 21-year-olds in October 1988.

Source: Special tabulations of the 1988 October and November CPS

Figure 2. Ratios of estimated dropout rates to CPS dropout rates noninstitutional population

Reporting category	Simple adjustment		Poststratified	
	Total dropout rate	Preceding school year dropout rate	Total dropout rate	Preceding school year dropout rate
Overall	76.7	79.6	85.1	83.1
Race				
White	76.1	77.5	84.7	80.4
Black	80.5	88.3	86.8	91.5
Other	84.7	76.6	92.9	87.5
Ethnicity				
Hispanic	76.2	63.5	87.4	68.8
NonHispanic	79.9	83.1	85.3	85.8
Sex				
Male	77.1	60.5	85.6	83.6
Female	76.3	77.1	84.6	81.1
Family income				
Less than \$10,000	71.2	75.8	77.8	76.4
\$10,000 - \$19,999	88.6	78.6	96.1	82.4
\$20,000 and above	96.8	100.0	103.3	102.0

SOURCE: Special tabulations of the 1988 October and November Current Population Survey

Figure 3. Ratios of estimated enrolled 3-5 year olds to CPS totals using simple and poststratified estimators, October 1988 noninstitutional population

Reporting category	Simple adjustment				Poststratified			
	Total	In school	In nursery school	In kindergarten	Total	In school	In nursery school	In kindergarten
Overall	100.0	103.5	106.3	101.0	100.0	101.6	102.6	100.1
Race								
White	102.8	106.9	108.0	104.0	100.1	101.2	101.8	100.5
Black	84.5	88.7	88.5	86.1	100.1	103.0	106.2	98.2
Other	98.5	106.5	111.6	98.7	98.8	103.3	108.6	98.4
Ethnicity								
Hispanic	87.0	92.1	98.8	88.3	100.0	105.6	118.1	100.0
NonHispanic	101.7	104.6	106.7	102.7	100.0	101.1	101.7	100.0
Sex								
Male	98.6	102.7	107.0	99.2	99.9	101.0	102.7	96.7
Female	100.4	104.3	105.7	103.0	100.2	102.3	102.5	101.5
Family income								
Less than \$10,000	72.8	78.5	83.0	75.2	82.9	87.8	96.1	83.0
\$10,000 - \$19,999	98.3	101.5	100.3	101.7	103.3	104.2	101.0	104.6
\$20,000 and above	112.0	112.4	112.4	112.3	106.5	105.1	104.4	106.6
Household type								
Husband and wife	105.2	108.0	110.1	106.2	103.1	104.1	103.8	103.7
Other family	82.2	86.4	88.6	83.7	89.3	91.8	96.7	87.5
Other type	97.1	—	—	—	100.3	—	—	—

SOURCE: Special tabulations of the 1988 October and November Current Population Survey

* The value of the estimate was suppressed because the base was less than 50,000

Session 4
TELEPHONE DATA COLLECTION

QUALITY IMPROVEMENT IN TELEPHONE SURVEYS

Leyla Mohadjer
David Morganstein
Westat, Inc.

1. Introduction

The use of telephone as an alternative mode of data collection in surveys has become very popular in recent years. Considerable research has been dedicated during the past decade to evaluate the quality of data collected in telephone surveys and to compare that with data collected by face-to-face interviewing. Simultaneous to the increased use of this methodology has been efforts at improving its efficiency and reducing the total error of telephone survey estimates. This paper provides a summary of recent methods for improving the quality of telephone surveys and reviews the recent literature on the results of these efforts.

Below we discuss several aspects of telephone surveys that fall into the category of "quality improvement." Most of these issues are design decisions that affect the expected total survey error. From its very beginning, the choice of telephone sampling over face-to-face sampling was one of improved efficiency. That is, the cost per complete in almost every case is significantly less than that of face-to-face sampling while the 'quality' of the results, as measured by total survey error, is little if any reduced. By way of comparison, mail-out surveys may have a very low cost per complete, but they suffer from large and generally unknown biases. Increasing efficiency is a traditional argument for system changes, such as the choice of telephone sampling over face-to-face interviewing, whose principal purpose is that of quality improvement.

In the following sections, we discuss several aspects of telephone survey operations in which the quality of a telephone sample design is established. We begin with decisions regarding the survey methodology. These decisions typically include the trade-off of greatly reduced survey cost for what might be, at most, a small increase in mean square error (MSE). Less quantifiable in cost terms is the reduced time to completion of survey operations afforded by a telephone methodology and improvements in the level of quality assurance.

Next, we discuss a number of sample design aspects which impact on the survey cost, schedule and error. We mention the much discussed issue of coverage and the general problem of frame construction as they relate to total error. A number of sample design improvements have been developed in the past few years which can decrease the expected number of wasted calls needed in the

process of identifying eligible respondents. These are described and compared.

As contrasted to other methodologies, telephone surveys contain a number of operational features which result in improved quality. Among these are aspects of management and supervision and of direct data entry through Computer Assisted Telephone Interviewing (CATI). We discuss and quantify some of the benefits which accrue from these approaches. Lastly, we review the topic of estimation as it relates to minimum mean square estimates. Several estimation procedures are required by the sample designs which are worthy of note.

2. Overview of the Properties of Telephone Surveys

Telephone surveys have become an often selected alternative to face-to-face interviewing for several reasons. Telephone surveys can be conducted at a much lower cost when compared to face-to-face interviewing. They also allow for the sample results to be available more quickly than face-to-face surveys. There are greater opportunities for quality control through more rigorous supervision and through frequent monitoring of the interviewing staff. Also, telephone interviewing makes it possible to contact otherwise hard-to-reach respondents such as those living in difficult to visit or dangerous neighborhoods, in bad weather conditions, or late at night (Groves and Kahn, 1979). The sample design effects for estimates derived from telephone surveys are smaller than those coming from more heavily clustered area probability designs. Finally, telephone surveys have smaller interviewer effects. Discussions on these issues are provided in different sections of this paper.

Considerable research has been dedicated to improving sampling techniques, to increasing response rates, and to reducing noncoverage bias. Research has also focused on the issue of data quality, a comparison of collection modes, and the influence of collection mode on the quality of the data. Several authors such as Groves (1979) and Jordan (1980) have stated that one of the causes of lower performance for telephone surveys when compared to face-to-face surveys is the lower degree of operational experience with telephone surveys. Leeuw and Zouwen (1988) have analyzed the results of a number of studies in this area. Their work confirmed that the difference between the face-to-face and telephone interviews is becoming smaller over time.

Leeuw and Zouwen (1988) integrated findings on interviewing mode differences and have provided a review of this topic. The method of analysis they used made it possible to present an overview of mode differences found with respect to data quality and estimate the size of these differences. The main conclusions of their paper are the following:

- Response rates are generally higher for face-to-face interviews than for telephone interviews;
- The majority of studies did not find statistically significant differences in modes. When differences were found, however, they were in favor of face-to-face interviews; and
- Only small differences were found between random digit dialing (RDD) and face-to-face, and the differences have become smaller over time.

Leeuw and Zouwen (1988) also point out that one major difference between the two modes is the lack of visual support in telephone surveys. This makes the respondent's task of answering some questions difficult in telephone surveys. It also results in reduced control over the respondent's behavior in telephone surveys. On the other hand, since the questions come through the phone, responses are meaningless for other persons in the same room with the respondent, especially for closed questions. This reduces the potential influence of "bystanders" on the respondents.

The fact that telephone interviewing can be contained in a small area offers many potential benefits. Interviews done by telephone are subject to more supervisory control than field surveys, resulting in a positive effect on the quality of data from telephone surveys. Unlike the face-to-face mode, supervisors can monitor telephone interviewing anonymously and frequently with little impact on survey costs. This allows for rapid modification of questionnaire wording found to be problematical. In addition, they can arrange for needed interviewer re-training or they can make appropriate re-assignments if an interviewer is observed to be unsuitable for their assignment. In addition, with CATI systems, it is much easier to put checks and probes in different parts of the interview to insure that answers provided by respondents are consistent throughout the questionnaires. All of these features should result in reduced non-sampling error.

Two disadvantages of telephone surveys are the noncoverage of persons living in households without telephones, and lower response rates when compared to face-to-face surveys. Section 3 provides a discussion of undercoverage in telephone surveys and the methods available to compensate for the undercoverage. Section 4 discusses nonresponse issues in telephone surveys.

3. Undercoverage in Telephone Surveys

Households without telephones are not included in telephone surveys since the sampling frames do not include such households. A considerable amount of information has been published on the nature of possible biases resulting from the use of a telephone sampling frame. Thornberry and Massey (1988) have analyzed trends

in telephone coverage in the U.S. across time and subgroups of the population. They indicate that estimates for the entire U.S. population may experience only minor biases because of the high rates of telephone usage, about 93 percent of the population can be reached by telephone.

Although overall telephone coverage has risen to a very high level, it is not uniformly distributed across the population. Thornberry and Massey (1988), Groves and Kahn (1979), and Banks (1983) have shown striking differences between telephone and non-telephone households with respect to demographics, economics, and health characteristics.

As might be expected, telephone coverage correlates highly with income. Massey (1988) points out that other variables such as employment status, education, marital status, and race are also correlated with income and thus affect telephone coverage. More lower-income persons tend to be missed in telephone screening. This, in effect, results in higher telephone penetration for whites than blacks. Telephone coverage is lower in the South than in the rest of the U.S., and it is lower in rural than urban areas.

Massey (1988) points out that noncoverage bias is a function of the noncoverage of a telephone survey frame, and of the difference in characteristics between the covered and uncovered population. Even though the percentage of households with telephones may increase and the overall noncoverage rate becomes smaller, large differences between telephone and nontelephone households can result in significant noncoverage bias. Surveys which focus on income or variables related to income may experience high noncoverage bias. It is true that the estimates of characteristics for the total population may not be drastically affected by the omission of nontelephone households, however, for some subdomain estimates there could be large biases due to the exclusion of households without telephones.

3.1 Methods to Compensate for Undercoverage

Several methods are available in telephone surveys to address the problem of noncoverage bias. One approach which may eliminate certain kinds of undercoverage bias is the use of dual frames. Dual frame, mixed mode surveys use a combination of RDD and face-to-face samples to overcome the noncoverage of households without telephones. Research in the area of such mixed mode surveys include Sirken and Casady (1988), Groves and Lepkowski (1985), Lepkowski and Groves (1984), Biemer (1983), and Casady et al. (1981).

Sample weighting adjustments in the form of post-stratification factors can be used to decrease the effects of noncoverage. The post-stratified weights are frequently employed

in national surveys to compensate for noncoverage bias. The subgroups established for the purpose of post-stratification are specifically tailored to each study. Subgroups are defined on the basis of variables thought to be correlated with the major statistics to be obtained from the survey, as well as variables correlated with telephone penetration and nonresponse distribution. Massey and Botman (1988) have investigated the impact of post-stratification survey adjustments in national surveys. They discuss several post-stratified weighting adjustment methods for RDD surveys, and show the effect of these adjustments on the estimates. Other work done in this area includes Banks (1983), Banks and Undersign (1982), and Thornberry and Massey (1978). Their results show that, although these methods reduce the effects of undercoverage, they do not completely eliminate the bias.

3.2 Within Household Coverage

The main focus of research in the area of coverage in random digit dialing surveys has been on sampling frame inadequacies, i.e., the exclusion of nontelephone households from the frame, as discussed earlier. However, there is another cause of undercoverage that arises from failure to obtain complete listings of household members in responding households. This is usually referred to as within household coverage. Within household coverage also exists in face-to-face surveys. Maklan and Waksberg (1988) used two surveys conducted by Westat and compared their within-household coverage rates with those obtained by the Current Population Survey (CPS). They concluded that the coverage of persons in households with telephones generally available in RDD surveys is at least as good, if not better, than that provided by CPS.

4. Nonresponse Issues in Telephone Surveys

Groves (1988) gives an overview of nonresponse issues in telephone surveys, and distinguishes between those factors common in both face-to-face and telephone surveys and those factors that are specifically related to the selection mode. Factors such as length of the questionnaire, subject matter (topic of the survey), sensitivity of the questions, refusal conversion and callback routines are common in both modalities. The differences in response rates that Groves (1988) cites between face-to-face and telephone surveys are that refusal rates are higher for telephone surveys, and relatively more of the refusals take place immediately after interviewers have introduced themselves prior to describing the purpose of the survey. However, as pointed out earlier, Leeuw and Zouwen (1988) have shown that these differences have become smaller over time. Researchers have varied the introductory section in an effort to reduce early refusals. A number of researchers have reported some improvements by using advance

letters to alert sample persons about the survey and the upcoming telephone call.

5. Choice of a Frame, List vs. RDD

Essentially, there are three types of sampling frames available for telephone surveys. List frames use information available in telephone directories, or other frames based on telephone directories, to generate telephone number sample. This is the alternative with the greatest undercoverage problem. Second, random digit dialing provides a frame of all possible telephone numbers, and thus covers both listed and unlisted numbers. Third is a multi-frame approach which uses both directories and RDD. Lepkowski (1988) provides a description of these frames and methods of sample selection used with them.

6. Two-Stage RDD

Random digit dialing was originally developed to overcome the coverage problems inherent in directory samples; however, surveys of residential respondents were burdened by the excessive effort required to filter many nonworking or business telephone numbers. The Mitofsky-Waksberg cluster sample technique eliminates much of this inefficiency by utilizing the manner in which the telephone industry initiates new phone exchanges which is to assign a prefix to either a business or a residential clientele. Accordingly, it is possible to select a probability sample that is significantly richer in residential numbers than would be obtained by conducting a simple random sample of telephone numbers.

6.1 Waksberg Method for Reducing Effort

The method frequently used for large scale residential telephone surveys is a two-stage cluster procedure. This method was originally developed by Mitofsky (1970) and Waksberg (1978), and is usually referred to as the Mitofsky-Waksberg method. In a 1978 article, Waksberg demonstrated mathematically that this procedure provides a probability sample of households with telephones, in which all telephone numbers have the same probability of selection. Further, the method was shown to require a smaller number of telephone calls than the sampling procedures previously used for RDD, and thus, as a quality improvement, significantly reduces the cost and time involved in such surveys in comparison with dialing numbers at random.

The majority of numbers dialed completely at random are nonworking, business and other nonresidential numbers. Current estimates are that about 75 percent of the potential numbers within

existing telephone prefixes are nonworking and another three percent are businesses or institutions of some type. Given that only about 20 percent are residential numbers, a typical RDD simple random sample requires that five calls be made to locate a single household. In some cases, the telephone companies do not provide a message that the number dialed is not a working number. Additional checking necessary to distinguish between not-at-homes and nonworking numbers adds further to the cost of achieving completed interviews.

The Mitofsky-Waksberg sampling method is designed to reduce the number of nonproductive calls. It takes advantage of the fact that a high proportion of nonworking and commercial numbers occur in consecutive sequences. Essentially, the procedure involves two steps: first, "household cluster identification" (identifying and selecting a sample of blocks of 100 numbers called "telephone clusters," which contain working, residential telephone numbers); and second, dialing random numbers within the clusters. Users of this technique typically locate three residential numbers for every five attempted within each cluster, a significant improvement in efficiency for minimal additional effort.

6.2 Modified Waksberg Method

The "standard" Mitofsky-Waksberg method, which produces a self-weighting sample, involves designating a desired number of household clusters, and sampling a constant number of households per cluster. There are, however, some awkward operational features arising from the requirement for a constant number of households per cluster. For example, before the need for more telephone numbers in specific clusters can be determined it is necessary to wait until the required number of households have been identified and interviewed. Since a large number of calls are required to determine whether a telephone number is residential and, if so, to obtain the cooperation of the household, the standard method is rather time-consuming.

To improve the data collection process and to reduce the data collection time, researchers have come up with different ways to speed up the data collection (for example, refer to Alexander [1988], Potthoff, JASA [1987], and Potthoff [1987]). The modified Waksberg procedure that Westat sometimes applies is based upon a fixed number of telephone numbers (instead of households) per cluster. There is thus no necessity to wait until the original sample of clusters has been completed to determine whether the desired number of households within clusters has been achieved. With the modified method, sample size becomes a random variable and the tight control on sample size offered by the original procedures is loosened. What is more, the modified procedure results in a sample that requires sample weights to adjust for differential probabilities of selection. Accordingly, its reduced data

collection time is purchased at the price of increased sampling error.

7. Efficiency of Estimates Derived from RDD Studies

In designing a two-stage RDD sample, the number of sample clusters and the average number of sample households per cluster must be specified. The choice of the sample sizes is usually made on the basis of cost and variance considerations. The extent to which the variances are increased due to clustering depends on the intraclass correlation between households within cluster and the average number of eligible households per cluster.

Clustering generally reduces survey data collection costs. The magnitude of the cost, however, is very different for face-to-face than it is for telephone surveys. The cost savings brought about by reduced travel costs is a virtual necessity in face-to-face surveys wherein they could comprise a substantial portion of the total survey cost. In telephone surveys, clustering is used to reduce the cost of dialing and reaching telephone numbers that belong to households. Considering the minimal cost of dialing telephone numbers (especially when compared to travel cost in face-to-face surveys), cluster sizes in telephone surveys need not be as large as they are in face-to-face surveys. As a result, statistics derived from RDD surveys are generally more efficient (have smaller variances) than those coming from face-to-face surveys.

8. Improvements in Locating Rare Populations in Telephone Surveys

Studies of specific subgroups of the population that comprise relatively small proportions of the total population have always been the focus of many research efforts. With any method of sample selection, surveys of rare populations almost always require a considerable amount of screening. The frame generally used for RDD, a computer file provided by AT&T, comprises all telephone households. Subsets cannot be determined except as part of a screening procedure. Extensive screening is necessary to locate members of the rare population, and as a result, it is usually very costly to sample rare populations through telephone surveys.

One efficient option for sampling members of rare populations is to use a commercially available tape (the Donnelley tape) that contains census population characteristics for prefix areas. Mohadjer (1988) provides an evaluation of the quality of the information on this tape. Furthermore, Mohadjer (1990) discusses the effectiveness of using the Donnelley tape to improve the sample efficiency in an education study. She shows that sampling efficiency is greatly improved by using the Donnelley tape to oversample blacks and Hispanics.

9. Interviewer Effects in Telephone Surveys

Many studies have compared interviewer effects in telephone and face-to-face surveys. They mainly speculate that the interviewer effect is smaller for telephone surveys than for face-to-face surveys. In this section we examine the potential causes for interviewer effects and the way these causes relate to the data collection mode.

Stokes and Yeh (1988) give the following as the potential causes for interviewer effect:

1. Not following directions exactly;
2. Variations in personalities, tone of voice;
3. Respondent's reaction to characteristics of the interviewers; and
4. Different response rates for different interviewers.

The main belief is that the variability among interviewers is smaller in centralized telephone surveys than in face-to-face surveys. The reason often given is that these effects can be controlled better by monitoring and supervision in a centralized data collection facility. Telephone interviewers can be much more easily monitored and training can be more uniform as well as more frequent. Furthermore, interviewers have the opportunity to observe and learn more from each other in a centralized facility such as a telephone center. This makes interviewer behavior more uniform in telephone surveys than in face-to-face surveys. Differences between interviewers can be detected much easier, especially in centralized facilities. When differences are observed between interviewers, steps can be taken readily to reduce them. For example, changes in training or instructions to interviewers can be implemented more quickly.

The interviewer's personality and the respondent's reactions to the interviewer also have smaller effects in telephone surveys. The tone of voice is the only variable that is thought to have a higher effect in telephone surveys than in face-to-face surveys. This effect is suggested because of the lack of visual contact in telephone surveys (the lack of visual contact increases the effect of tone of voice on respondents).

A number of steps can be taken to limit these interviewer effects even further. There are several quality control measures which can provide a quick assessment of interviewer performance and which can identify the need for action. Strict supervision is especially important in the early stages of data collection to

insure that all interviewers are following directions and have a clear understanding of the survey purpose and instrument. Group meetings to emphasize important aspects of the procedures and individual conferences with weaker interviewers should be used to limit the effect of interviewer differences. All interviewers should be monitored when they first begin data collection. Staff who fail to meet or exceed standards should not be allowed to continue until they have undergone remedial training.

10. Computer-Assisted Telephone Interviewing (CATI)

The use of CATI was a quantum step in telephone survey quality improvement. Survey organizations have used CATI with increased frequency in recent years because of its many benefits. It is believed that CATI improves the quality of the data collected, it reduces the cost of data collection, and it increases the timeliness of telephone surveys.

A CATI system has the potential for providing clean data immediately after interview completion. Three CATI features contribute to this capability. First, most data edits can be done on-line as the responses are entered. A CATI program can prevent interviewers from entering out-of-range responses ("hard range check") and can be programmed to require verification of unlikely responses ("soft range check"), e.g., such as an age of 100 years. Second, consistency checks are possible in the CATI program appropriate to inconsistent responses verified during the interview. Third, CATI can be set up so as to prevent the interviewer from leaving a question incomplete. If the interviewer has difficulty recording an answer (e.g., difficulty categorizing the answer into the precoded choice on the CATI screen), they can be trained to enter a "comment" explaining the circumstances. A quality control monitor can be responsible for reviewing all interviewer's comments on a daily basis to resolve difficulties and to update the data files as required.

It was previously observed that a telephone interview methodology helps to reduce between-interviewer variances because of greater opportunity for monitoring and supervision. Since interviewers can be easily observed without disturbing the interview process, frequent monitoring can be used to uncover interpretation and presentation difficulties, all of which contributes to reduced interviewer variance. A CATI approach represents yet another step in this same direction. Between-interviewer differences in understanding the flow of the instrument can be virtually eliminated.

Often sampling efficiency can be improved through the use of complex respondent selection procedures. Unfortunately, complexity can breed errors especially when interviewers are tired or are dealing with a difficult set of questions. Through the use of

CATI, very complicated sampling rules can be implemented, virtually without error. The interviewer enters the household composition and the software selects a random respondent using pre-specified sampling rates.

As pointed out by Nicholls (1988), additional advantages of the CATI systems can be summarized in the following way:

- Rather than being managed by the interviewers, the status of each sampled case is available in the computer, thereby improving sample management.
- The scheduling and assignment of cases are done by computer. The scheduler schedules the appropriate time to call respondents taking into account the time differences across the U.S.
- On-line interviewing makes it possible to display the instruction to the interviewers, display the survey questions, and response categories without any need to use paper and pencil.
- Answers to closed questions which are not in the permissible range can be determined at the instant the response is entered. The software can prompt the interviewer that this answer contradicts another response given by the same respondent at an earlier point in the interview and a correction can be made immediately. This reduces the need for data retrieval.
- Branching or skipping to the next item is done by the computer. This improves the quality of data collected for more complex data collections that involve complicated skip patterns and subsampling at different stages of data collection.
- Interviewers may interrupt, resume or repeat some of the sections. Also they can go back and correct previous answers or write notes on the screen in appropriate places.
- The system improves supervision. The screen and the telephone conversation can be seen and heard with no disturbance to the interviewing process. The telephone conversation with the respondent can be monitored. All of these advantages result in faster reaction to the needs for clarification, re-training or re-assignment.
- The survey results are virtually ready for weighting and tabulation upon completion of the data collection phase. This more timely data collection makes possible survey schedules that could not have been met in the past.Ñ

- The CATI system maintains records of on-line calls, outcomes of the calls, response rates, and the amount of time spent by interviewers. It can also be used to time different parts of the questionnaire if the survey length becomes a problem.

Since effective CATI interviewers must be able to perform a number of demanding tasks simultaneously, the task of training suitable staff is more challenging. Interviewers must establish rapport with a respondent, accurately read the questions shown on the terminal screen, correctly code the response, and enter messages to the respondent's file indicating that a probe (e.g., reading a question, prescribed clarification of an item, etc.) was required. In addition, they must record verbatim a respondent's comments on a question, and keep the respondent's interest long enough to complete the interview. This is a set of qualifications that require interpersonal, computer, and typing skills that surpass those of traditional telephone interviews. Fortunately, the improved ability to monitor telephone interviews conducted via CATI assist in assuring that suitable staff is adequately prepared for the survey.

11. Summary

Face-to-face interviewing has long been the standard data collection method selected when the highest quality survey results were required. The authors have reviewed those features of telephone surveys which can result in improved survey quality, that is, reduced total survey error for the same, or even for less, cost as other modes such as face-to-face interviewing. After review of these features, survey designers are better able to choose between a telephone sampling approach and a face-to-face methodology.

References

Alexander, C.H., "Cut Off Rules for Secondary Calling in a Random Digit Dialing Survey," Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988.

Banks, M.J., "Comparing Health and Medical Care Estimates of the Phone and Nonphone Populations," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1983, page 569-574.

Banks, M.J., and Anderson, R.M., "Estimating and Adjusting for Nonphone Coverage Bias Using Center for Health Administration Studies Data," in National Center for Health Services Research, Health Survey Research Methods: Proceeding of the 4th Biannual Conference, National Center for Health Services Research Proceeding

Series, Department of Health and Human Services Publication No. (PHS) 84-3346, 1982.

Groves, R.M., and Lyberg, L.E., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 191.

Groves, R.M., and Kahn, R.L., Surveys by Telephone: A National Comparison With Personal Interviews, New York, Academic Press, 1979.

Jordan, L.A., Marcus, A.C., and Reeder, L.G., "Response Styles in Telephone and Household Interviewing: A Field Experiment," Public Opinion Quarterly, Vol. 44, No. 2, Summer 1980, page 210-222.

DeLeeuw, E.D. and Van der Zouwen, J., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 283.

Lepkowski, J.M., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 73.

Maklan, D., and Waksberg, J., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 51.

Massey, J. T., and Botman, S.L., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 143.

Massey, J. T., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 3.

Mitofsky, W., "Sampling of Telephone Households," unpublished CBS memorandum, 1970.

Mohadjer, L., "A Study of the Effectiveness of Oversampling Telephone Clusters with High Concentrations of Blacks and Hispanics in the NHES Field Test," unpublished report to NCES, 1990.

Mohadjer, L., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 161.

Nicholls, W.L., II, Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 377.

Potthoff, R.F., "Some Generalizations of the Mitofsky-Waksberg Technique for Random Digit Dialing," Journal of the American Statistical Association, Vol. 82, No. 398, June 1987, pp. 409-418.

Potthoff, R.F., "Generalizations of the Mitofsky-Waksberg Technique for Random Digit Dialing: Some Added Topics," Proceedings of the

Section on Survey Research Methods, American Statistical Association, 1987, pp. 615-620.

Sirken, M.G., and Casady, R.J., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 175.

Stokes, L., and Yeh, M., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 357.

Thornberry, O.T., Jr., and Massey J.T., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 25.

Thornberry, O.T., Jr., and Massey J.T., "Correcting for Undercoverage Bias in Random Digit Dialed National Health Surveys," Proceeding of the Section on Survey Research Methods, American Statistical Association, 1978, page 224-229.

Waksberg, J., "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, Vol. 73, No. 361, March 1978, page 40-46.

COMPUTER ASSISTED SURVEY TECHNOLOGIES IN GOVERNMENT: AN OVERVIEW

Marc Tosiano
National Agricultural Statistics Service

Introduction

CATI is an acronym for computer assisted telephone interviewing. It is the interactive use of computers to assist in data collection activities typically performed in a centralized telephone facility of a survey organization.^{27 31} CATI is only one use of the computer in the growing realm of computer assisted survey work. Other uses of computer assisted surveys include: 1) computer assisted personal interviewing (CAPI), 2) computerized self administered questionnaires (CSAQ), 3) computer assisted data entry (CADE) of information on paper questionnaires into a electronic format.^{31 26} Each of these computer assisted survey techniques may be used alone for a survey or in combination depending on survey management requirements and the various modes used to collect data for a given survey.

Features of computer assisted surveys

During an interview, the minimum use of computer assisted survey technology is the presentation of survey questions and their response categories on the computer screen. Interviewers read the question to the respondent and key the answer on the screen by using the computer keyboard. However, computer assisted survey techniques offer many capabilities above and beyond the traditional paper questionnaire. These features include enhancements to the interview proper as well as the automation of survey management activities. Obviously, the features available depend on the software chosen for computer assisted interviewing.¹⁰ Common features offered by various computer assisted survey software include: ^{26 10 24}

On-line interviewing:

- * Instructional or reference information appears on the screen or is available via help screens to assist the interviewer.
- * Fills are used to customize question wording by inserting input from records prior to the survey or from answers to previous questions.

- * Answers to closed questions are checked against permissible entries. Some software offers multiple responses as well.
- * Numeric answers are checked against a pre-defined range.
- * Consistency checks are made against data collected earlier in the interview.
- * Answers detected as invalid can invoke an error correction routine or additional probing questions.
- * Formats are available for special answers, e.g., date, time, money, zip code, etc.
- * Open-ended questions or interviewer notes are answered by typing text.
- * Question order as well as response categories may be randomized to reduce order effects.
- * Item-based design offers one question per screen or multiple related questions per screen; the interviewer is forced to answer the questions in a pre-determined sequence.
- * Form-based design presents a screen that simulates a paper form. The interviewer is free to move the cursor around the form and fill in the form in any order.
- * Automatic branching is done based on input from records prior to the survey, previous answers in the interview, logical conditions, or arithmetic checks.

Creating the computer assisted questionnaire:

- * Some packages offer a menu driven approach to building the questionnaire while others require the use of a special programming language.
- * Some packages come with their own editor to write or change the questionnaire, but other editors or word processors may be used as well.
- * Questionnaire debugging tools of various strengths may be available.
- * A paper copy of the questionnaire including screen prints and a flow chart of the questionnaire may be available.

Survey management:

- * The sample is stored in computer media and the software maintains the status of each questionnaire.
- * Sampling procedures may be available including random digit dialing facilities.
- * Call scheduling delivers the next case to be called by the CATI interviewer. The call scheduler prioritizes and sequences the calls made in the CATI environment. This includes the retrieving of cases at the appointed time for a call-back, establishing follow-up calls for busy signals or no answers, and targeting groups of cases such as strata or replicates.
- * Survey managers may generate reports including such things as: completions, response rates, refusal rates, time per interview or question, call-back appointments, etc. These reports may be by interviewer, by day, by shift, cumulative, etc.
- * Monitoring individual CATI interviews may be done by viewing the interviewer's screen at a supervisor's workstation where audio monitoring may be available as well.

Data handling and analysis:

- * Post-survey processing may be done to review, edit, clean, or code each interview.
- * A codebook may be created containing questions, the variable names and location in the dataset, etc.
- * An audit trail may be maintained with all previous answers if an answer is changed.
- * Output files are created in a form ready for the next processing stage, these could include SPSS and SAS datasets.
- * Some packages offer their own statistical analysis packages, including histograms, distributions, regression, Analysis of Variance (ANOVA), etc.

The features listed above are not available in all computer assisted survey software. A survey organization procuring software for a computer assisted application would have to decide which features are important and select software accordingly. In

addition, all software packages will not operate on all computer hardware; a problem for all computer systems which must be resolved is the matching of software to hardware.

Computer assisted survey software is relatively new and constantly evolving; enhancements are usually inspired by the needs and requirements of end users.¹⁰ Therefore, another consideration in choosing software might be the existence of a user support group and the willingness of the software company to enhance the system as new features are requested by users and the cost of these modifications.

These different features of computer assisted survey software have various effects on costs and quality of data. For example, the use of interactive interviewing may improve the quality of data, but without call scheduling, the productivity of interviewers may be unaffected.²⁶ If improved interviewer efficiency and the elimination of paper callback records is important, software with call scheduling would be more attractive. However, systems with call scheduling may not be strong in other areas such as form-based design or software portability across various hardware. Evaluating these trade-offs is a difficult but critical task in choosing (or developing) this type of software.

Costs and Data Quality

The CATI concept was originally proposed by the American Telephone & Telegraph Company; in 1971 they sponsored the first CATI survey to measure customer satisfaction.²⁶ After this experience, CATI was believed to have three advantages over conventional data collection methods: "accuracy, speed and reduced costs".²² Since then there have been many studies and papers evaluating the accuracy or extent of the validity of these original beliefs.^{26 22 9 19 35 41 37 16 17 18 33 36 38} Some authors have also reviewed the impact of CATI on survey administration and the internal structure of survey organizations.^{5 6 13 21 30 32} This section of the paper does not intend to review all of these sources but to briefly review some of the implications and consequences that arise by using this new computer assisted survey technology. Some of the topics discussed here are not easily definable as advantages or disadvantages; it often depends on the methods used to implement this new technology.

The first set of reasons for implementing computer assisted surveys is to expedite surveys and thereby reduce costs.²⁴ There is always the initial cost of procuring and maintaining hardware and software. This overhead cost could be alleviated by utilizing the hardware and/or software for projects other than computer assisted surveys.³⁵ Some of the hardware configurations used in the past have been 'dumb' terminals attached to a centralized mainframe

computer.¹⁴ Later, terminals or 'intelligent' microcomputers were attached to a minicomputer.^{27 35} The latest hardware innovation used is microcomputers used in a stand-alone or in a Local Area Network (LAN) environment.⁴ After these items are procured, there are the costs for training the staff to implement the new technology, and training the interviewers on use of the system. Interviewer training costs also depend on the turnover rate of interviewers. CATI questionnaire design will take longer than paper design because it employs many of the features listed previously such as automatic branching, use of fills, interactive editing and consistency checks, interviewer 'helps' and special processing needed for other activities previously done on paper.²⁰ This special processing includes resolution of busys, no answers, refusals, arranging callbacks and other administrative activities. As with other programming, CATI questionnaire designers typically 'steal' code from previous studies whenever possible. This efficient use of previous code is enhanced by the use of modular or structured programming. The CATI questionnaire setup for some surveys could be faster and simpler than creating a paper questionnaire, but only if the CATI instrument emulates the paper which is seldom the case.

Once past these overhead costs, there are other cost considerations. Interviews typically take longer with computer assisted surveys because of the edit checks and additional questions generated to probe for corrections or clarifications; another reason could be the interviewer's lack of familiarity with the keyboard, especially if there is a lot of text to be entered. These higher costs are somewhat offset by other features of computer assisted surveys. The use of an automatic scheduler can improve interviewer efficiency and reduce the cost of supervision by eliminating voluminous and tedious paper shuffling; supervisors are freed to do more real supervising rather than managing callbacks.⁵⁰ Status systems automatically keep track of each case in the sample including its current disposition and any actions taken on the case. Immediately after each interview, the data is already in electronic medium; this eliminates the data entry stage necessary in conventional data collection. At any time during the survey, output files are available for preliminary analysis and/or administrative reports needed to allocate resources during the remainder of the survey period.

The second and probably the more important set of reasons to implement computer assisted surveys is to improve survey data quality and enhance the ability to implement complex surveys.^{24 9} One major source of improved data quality is the ability to perform on-line edit and consistency checks which means corrections can be made during the interview with the help of the respondent. Post-survey edit checks can be eliminated or greatly reduced.³⁵ Many times, post-survey corrections to the interview are done without re-contacting the respondent; this results in more unknown or

imputed data. Computer assisted surveys also result in increased standardization among interviewers, especially in a central telephone facility.³⁵ ¹³ This standardization may help reduce some interviewer effects typically seen in paper questionnaires such as following proper question sequence.¹⁶ However, there are sources of error possible which did not exist in the paper environment such as simply keying the an incorrect number for an answer while using touch typing. Some of the benefits of complex instruments include: creation of multiple versions of a questionnaire within the same instrument, inclusion of pre-programmed probes, use of historic data from previous surveys, table look-up routines, and other techniques difficult to employ in a paper questionnaire. In addition, computer assisted technology permits easier implementation of research than does its paper counterpart. Some examples are: randomizing questions and answer categories, use of historic data, use of randomized probes to check respondents understanding of questions³⁰, re-interview and reconciliation studies, and item-based versus form-based questionnaire design.

Government CATI Implementations³⁹

Early CATI systems were developed by United States market research organizations in the late 1960's and early 1970's.¹⁴ University survey research centers became involved in this technology in the middle 1970's.²⁷ U.S. government agencies did not begin work with CATI until 1980 when both the Census Bureau and the National Agricultural Statistics Service (NASS) each established working groups to investigate this technology.² ²⁵ ³⁵

The largest installations of CATI in the federal government are in operation in four agencies: Bureau of Labor Statistics (BLS), Census Bureau, National Agricultural Statistics Service (NASS), National Centers for Disease Control (CDC). BLS has about 70 workstations in 14 sites. This includes a 10 workstation test site for developing CATI methods for the Consumer Price Index (CPI) Surveys which is planned for expansion to a 50 workstation production facility by 1994. Another 20 workstation site is in BLS headquarters for special surveys of the BLS Office of Employment and Unemployment Statistics. Their largest use of CATI is 40 workstations in 12 sites for the monthly establishment survey supporting the Current Employment Statistics Program. CATI is used for interviewing, non-response follow-up, and failed edit reconciliation. If successful, BLS plans expansion of these 12 sites to all 51 State offices with about 200 workstations in 1994.

The Census Bureau has two CATI sites with about 100 workstations. One site of 30 workstations is the Field Division's Hagerstown Telephone Center which collects data for surveys of household residents and small surveys of industry. This site is expected to expand from 30 to between 250 and 300 workstations by

1994. The second Census CATI site is in Jeffersonville, Indiana where 70 workstations are used to collect data from establishments for the Retail and Wholesale Trade Industries. Here, CATI is used for telephone interviews, data capture from paper questionnaires, and failed edit follow-up.

The National Agricultural Statistics Service (NASS) surveys farm operators and agricultural businesses with the largest CATI network in the Federal government. NASS has about 200 workstations operational in 14 State offices. Four additional State offices have recently installed the hardware and software for CATI and will soon become operational. This brings the NASS CATI capabilities to about 260 workstations in 18 State offices. Current plans are to install Local Area Networks in 42 State offices by 1992; this will increase the CATI workstation count to about 750 nationwide. While mostly used by CATI interviewers after business hours, these same workstations will also be used by the office staff during the day for normal office operations. These daytime operations include survey activities (e.g., data capture of paper questionnaires, interactive error detection and correction of data collected, survey management) and all other office work (e.g., word processing, spreadsheet operations, graphics).

The National Centers for Disease Control (CDC) operates about 150 workstations in 21 State offices to collect data for the Behavioral Risk Factors Surveillance Survey and other random digit dialed household surveys. Little expansion of the CDC CATI network is expected over the next few years because data collection is commonly contracted out to other survey organizations.

These Federal agencies are expanding their CATI capabilities and plan to complete their initial CATI implementation by 1994. Unlike many private and university survey organizations, government CATI installations are not generally implemented in a national or regional centralized telephone facility. Most of the federal resources are directed toward smaller State offices where the same equipment is used for other survey related activities and office automation (BLS, CDC, NASS). Even with this increase in CATI activities, CATI will not become the only mode of data collection. Mailed questionnaires are still important in the mixed mode method of data collection in NASS and BLS. Personal interviewing is still important to all agencies as well, often as part of mixed mode data collection; the field interviewing staff numbers about 3,000 in the Census Bureau and about 2,800 in the National Agricultural Statistics Service. With these large field staffs, implementing CAPI may be the next large task facing computer assisted survey work in these agencies.

In private and university survey organizations, the use of CATI is generally associated with a single centralized telephone facility. CATI encourages a centralized facility to benefit from some of the features listed earlier such as automatic call

scheduling, monitoring, and administrative reports.^{13 40 1} A central facility is better suited to computer assisted operations because of the shared hardware, software, sample, and technical support. While CATI improves standardized interviewing and quality control (by automatic branching, tailored question wording, and probes for on-line edits), centralization contributes to survey management with consolidated and more standardized training and supervision of interviewers. One disadvantages of centralization may be that the interviewers do not have the local knowledge, and cultural understanding which local interviewers may share with the respondents.²¹

A major challenge to federal agencies implementing CATI involves the resolution of the associated issue of centralized or decentralized interviewing. Many agencies already have national, regional, and/or State offices with commitments to Federal-State agreements, office staff, and an interviewer staff including office and field interviewers. These commitments may have as much impact on implementation decisions as the goals of operational efficiency and maximizing data quality. The Census Bureau has transferred the Retail and Wholesale Trade survey from the traditional regional telephone calling to one centralized CATI facility in Indiana. The other previously mentioned three agencies have maintained their dispersed data collection techniques by implementing CATI in the existing regional or State offices. However, these dispersed CATI facilities can be used as central sites as well. For example, if a given sample is so widespread across the country, one or more State offices can be designated as regional CATI centers for that survey.⁵ NASS has successfully tested the centralization of CATI interviewing in regional centers while personal interviews were still administered from the State offices. However, this mixed mode with centralization for only part of the data collection requires strong communication, coordination, and overall survey management.⁵

Other organizational considerations revolve around the question, "How do computer assisted techniques fit in with the current mode of operations?" Some of these considerations may be specific to a survey or addressed for overall computer assisted operations. A few examples follow: What is the role of the supervisory interviewer? Should CATI edit checks during the interview or during post survey processing totally replace existing batch edits? How should technological advances in software and hardware be incorporated into an existing CATI operation? When a mailed questionnaire is followed up with CATI or CAPI, how closely should the interview instrument follow the paper questionnaire?¹²

In many cases, the difficulties of implementing computer assisted techniques in government agencies arise from organizational requirements, not the technology itself. Some of the problems encountered with CATI are due to use of a central

facility; these problems would be the same if paper questionnaires were used in the same central environment. Therefore, it is important to understand the source of potential problems when advocating or implementing a computer assisted system -- technological and organizational.

The Future of Computer Assisted Technology³

As these four government agencies are approaching full CATI implementation, newer technologies are developing which go beyond telephone interviews and some re-evaluation is necessary. Very little research has been done to measure the cost, timeliness, and data quality of surveys done with these new approaches. This paper reviews some of the major new technologies and their possible use by survey organizations. These technologies can be divided into five groups: computer assisted personal interviews, computer assisted self administered questionnaires, geographic and communication technologies, voice technology, and artificial intelligence.

Computer Assisted Personal Interviews

Now that computers are getting smaller and smaller, computer assisted personal interviewing (CAPI) is the next natural extension of computer assisted interviewing beyond CATI. As mentioned before, personal interviewing is still important in federal survey agencies; CAPI can be used to benefit from the advantages listed earlier and also improve the data transfer between personal and telephone interviewing for mixed mode surveys. Unlike the course of CATI development, government agencies are in the forefront of CAPI development both for their own use and in sponsoring CAPI investigations by universities and the private sector. Also, the government's implementation of CAPI is proceeding rapidly compared to CATI. CAPI investigations have found that CAPI data collection is acceptable to most respondents and that most experienced field interviewers can be trained in its use.^{7 23 42 44 3 43 49}

In addition to the organizational considerations of implementation of CAPI there are some technological problems which need to be addressed. Assignments and questionnaires must be given to CAPI interviewers and completed interview data must be sent back to the office. National Analysts have used the mail, UPS, and courier services for this transmittal during the Nationwide Food Consumption Survey.⁴⁴ Another method is to use automated telecommunications with modems attached to computers. Research Triangle Institute (contracted by the Environmental Protection Agency) and the Netherlands Central Bureau of Statistics have used telecommunications with some success.^{43 49} However, the Netherlands is returning to the use of mail for data transmission as a simpler

and less costly approach.⁴⁶ If a workable solution is found, rapid telecommunications between the office and interviewers may be especially advantageous when operating on tight deadlines and using mixed mode methods. The Census Bureau and NASS plan to investigate an integrated CATI-CAPI system where cases can be transmitted between CATI interviewers in central or state offices and CAPI interviewers dispersed throughout the field.

The software used in CAPI is typically the same as used for CATI or personal interviewing in an office environment. However, personal interviews in the respondents home or at the doorstep can be more demanding and distracting. This may call for special software features for question formats, entry modes and questionnaire movement commands which are easier to use. These features specific to CAPI interviewers have not yet been determined or shared.

The hardware used for CAPI applications is still evolving and being investigated. Machines must also be evaluated based on the environment expected for conducting interviews: on a table top, standing and holding the machine, or both. The machines generally available for CAPI include laptop computers, hand held computers, and slate computers (handwritten character recognition devices). The laptops are generally 4 to 15 pounds and have various sized and types of screens and keyboards. Hand held computers are much smaller but offer very small screens, keyboards, and limited computing power which eliminates some software packages. Slate computers range from 3 to 4 pounds and are held like a clipboard while the interviewer reads questions and writes the answers on the screen with a stylus. This device emulates paper questionnaire data entry and some machines are able to recognize special functions such as tallies, diagrams, maps, and signatures. Unlike a year ago, these devices now run DOS based systems and NASS can run both BLAISE and CASES computer assisted software for CAPI applications on the Gridpad machine.

The weight of these machines is an important factor in an interviewer's acceptance of using a machine as a data collection tool. Most recent tests of CAPI have been qualitative reports with inconsistent findings.^{42 3 43} However, recent laboratory research has studied ergonomic properties of CAPI, interviewer attitudes, and logistical features of the technology. This work investigated the maximum weight of laptop computers which would lead to the acceptance of CAPI by interviewers for doorstep interviewing; further research is being done to include newer lighter laptops and slate computers.⁴⁵

Once the technological problems are resolved, survey organization and management will require review and modification to meet the needs of a computer assisted survey environment. CAPI may change the methods of assigning, conducting, supervising, checking-

in, and reviewing interviews. These changes will affect staffing requirements and how to most effectively organize and manage survey personnel. For example: 1) CAPI field supervisors must cope with hardware, software, and telecommunications problems in addition to interpersonal skills. 2) Interviewer training must include machine maintenance, CAPI interviewing, and transmission of assignments and data. To reduce costs, some of this training could be done as home study with on-line tutorials. 3) The software will eliminate survey specific errors such as inappropriate skips or data inconsistencies; however, supervisors will need to identify interviewers needing further training in CAPI operations. 4) Field supervisors and office staff must use new techniques to check-in, review, and edit CAPI interviews. Due to computerization, some of these functions may also disappear requiring clerical staff to be replaced with technical staff. 5) With better communications and data transmission, the relationship of State, regional, and headquarters staff may change as well. Data and messages could travel directly between field interviewers and headquarters. All these possibilities and more will affect how CAPI is implemented in the various survey agencies.

Computerized Self-Administered Questionnaires

Establishment surveys usually collect brief numeric responses from the same respondents time after time. New technologies may be welcomed by these respondents if it results in reduced respondent burden or is perceived as such. This area is ripe for the investigation of computerized self-administered questionnaires (CSAQ).

BLS is experimenting with voice simulation of the questions and touchtone data entry of the answers by the respondent.^{31 29 4} When respondents have prepared their reports, they dial a local telephone number at a nearby BLS office; a voice simulation module requests the entry of their identification number on the telephone's touchtone pad. The voice module then asks survey questions that the respondent answers by keying the numeric response on the touchtone pad. Since this procedure operates 24 hours, this interaction can be done at the respondent's convenience; without a telephone interviewer and data entry staff, costs are minimal. Of course, a BLS interviewer is still needed to call non-respondents after a cutoff date or to resolve data inconsistencies. A further extension of this project is voice recognition of the respondent which would eliminate the need to key answers on the touchtone pad.⁴⁴

The Energy Information Agency (EIA) is investigating CSAQ by using respondents' personal computers.³⁴ Respondents who have access to personal computers are given diskettes containing the monthly CSAQ, menu-driven procedures to obtain the necessary

information from other files, and programmed procedures to electronically transmit the completed questionnaire to the EIA computer.

Geographic and Communication Technologies

Other technological developments may assist field interviewers in some of the administrative work accompanying personal interviews. These include automobile telephones, beepers, and navigational and position-recognizing systems to provide reliable geographic coordinates. This technology could be used to: 1) assist rural field interviewers in locating sampling units by using coordinate position of landmarks and buildings; 2) update maps by driving through new streets not on current maps; 3) define coordinates of area frame boundaries for sampling because these coordinates are not affected by changes in physical boundaries or political borders; 4) recording precise locations of dwellings and establishments to allow summation of data to any area definable by geographic coordinates. On recent examination, the Census Bureau found that current systems are not sufficiently accurate, reliable and cost-effective for typical survey applications.³⁹

Voice Technology

The National Bureau of Standards has recommended that this technology be investigated by the Census Bureau as the next step in computer assisted methods.²⁴ This technology includes both voice simulation and speech recognition. It could be used to conduct telephone interviews without human interviewers or as an auxiliary computer tool to reduce the keyboard skill necessary for interviewers using computer assisted methods. As mentioned earlier, some voice technology is being investigated for gathering data from establishments at their convenience. For household or other personal surveys, acceptance of a fully automated computer interview seems to depend upon respondent acceptance. However, the potential cost savings possible from voice technology will probably stimulate further research in this area; survey agencies will need to evaluate these new systems as they become available to judge their applicability to surveys.

Artificial Intelligence

Artificial Intelligence is a computer discipline which builds computer programs that perform tasks requiring intelligence when done by humans. This discipline is used to develop expert systems for problem solving which involve the use of appropriate information acquired previously from human experts.¹¹ This technology has been used by Westat for computer assisted coding of

open-ended questions on a paper questionnaire. Initially, humans do all the coding which is recorded by the computer and from this human input, the computer "learns" how to do this coding as well. As the coding process continues, the computer program can code increasingly more open-ended responses while the human operator can verify these codes and handles the responses not yet "learned" by the program.⁴⁷ Although this technology may have limited use during data collection, this may be a computer assisted technique which could benefit other survey management tasks like case assignments, questionnaire coding, and automatic call scheduling.

Conclusions

Government survey agencies have taken about 20 years to implement one new technology, CATI. Meanwhile, technology has advanced into many other areas such as computer assisted personal interviews, computerized self-administered questionnaires, geographic and communication technologies, voice technology, and artificial intelligence. This technology explosion means that survey agencies need to evaluate an ever increasing number of methods which may improve data collection and survey management. In addition to investigating new technologies, the associated organizational and methodological factors must be addressed so that all implications are considered before implementing advanced computer assisted survey methods. All the while, studies must continue to evaluate the effects of these factors on survey costs, timeliness, and data quality.

Acknowledgement

A note of gratitude goes to Bill Nicholls for material presented in his report as referenced in [39].

References

- ¹ American Statistical Association, Proceedings of the Section on Survey Research Methods, Washington, D.C.: American Statistical Association, 1978. Experiences with CATI in a Large-Scale Survey, by William L. Nicholls II, pp. 9-17.
- ² American Statistical Association, Proceedings of the Section on Survey Research Methods, Washington, D.C.: American Statistical Association, 1983. Measuring CATI Effects on Numerical Data, by Carol C. House and Betsy Morton, pp. 135-138.
- ³ American Statistical Association, Proceedings of the Section on Survey Research Methods, Washington, D.C.: American Statistical

Association, 1988. Development of a Computer Assisted Personal Interview For the National Health Interview Survey, by Stewart C. Rice Jr., Robert A. Wright and Ben Rowe.

⁴ American Statistical Association, Proceedings of the Section on Survey Research Methods, Washington, D.C.: American Statistical Association, 1989. Developing a Cost Model for Alternative Data Collection Methods: Mail, CATI, and TDE, by Richard Clayton and Louis Harrell.

⁵ Bass, Robert T., and Robert D. Tortora, "A Comparison of Centralized CATI Facilities for An Agricultural Labor Survey," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

⁶ Berry, Sandra H. and Diane O'Rourke, "Administrative Designs for Centralized Telephone Survey Centers: Implications of the Transfer to CATI," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

⁷ Birkett, N. J., "Computer-Aided Personal Interviewing: A New Technique for Data Collection in Epidemiologic Surveys," American Journal of Epidemiology, Vol. 127, No. 3, 1988, pp. 684-690.

⁸ Carpenter, Edwin H., "Software Tools for Data Collection: Micro-Assisted Interviewing," Social Science Computer Review, Fall 1988.

⁹ Catlin, Gary and Susan Ingram, "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Pater Methods in Centralized Interviewing," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

¹⁰ deBie, Steven E., Inede A. L. Stoop and Katrinus L. M. deVries, CAI Software -- An Evaluation of Software for Computer Assisted Interviewing, Amsterdam, Association of Social Research Institutes, 1989.

¹¹ Dictionary of Computing, New York, Oxford University Press, 1986.

¹² Dillman, Don A. Mail and Telephone Surveys -- The Total Design Method. New York: John Wiley & Sons, 1978.

¹³ Dillman, Don A. and John Tarnai, "Administrative Issues in Mixed Mode Surveys," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

¹⁴ Fink, James C., "CATI's First Decade: The Chilton Experience," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 153-168.

¹⁵ Freeman, Howard E., "Research Opportunities Related to CATI," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 143-152.

- ¹⁶ Groves, Robert M. and Nancy A. Mathiowetz, "Computer Assisted Telephone Interviewing: Effects on Interviewers and Respondents," Public Opinion Quarterly, Vol. 48, 1984, pp. 356-369.
- ¹⁷ Groves, Robert M. and Lou J. Magilavy, "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys," Public Opinion Quarterly, Vol. 50, 1984, pp. 251-266.
- ¹⁸ Groves, Robert M., "Implications of CATI," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 199-215.
- ¹⁹ Groves, Robert M. and William L. Nicholls II, "The Status of Computer-Assisted Telephone Interviewing: Part II -- Data Quality Issues," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 117-134.
- ²⁰ House, Carol C., "Questionnaire Design with Computer-Assisted Telephone Interviewing," Journal of Official Statistics, Vol. 1, No. 2, 1985, pp. 209-219.
- ²¹ Lyberg, Lars, "Administration of Telephone Surveys," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.
- ²² Nelson, R. O., B. L. Peyton and B. Z. Bortner, "Use of an On-Line Interactive System: Its Effects on the Speed, Accuracy, and Costs of Survey Results." Presented at the 18th Advertising Research Foundation Conference, New York City, 1972.
- ²³ Netherlands Central Bureau of Statistics, Automation in Survey Processing, Select Report 4, Voorburg, Netherlands, Central Bureau of Statistics, 1987.
- ²⁴ Nicholls II, W. L., "Computer-Assisted Telephone Interviewing: A General Introduction," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.
- ²⁵ Nicholls II, William L., "CATI Research and Development at the Census Bureau," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 191-197.
- ²⁶ Nicholls II, William L. and R. M. Groves, "The Status of Computer-Assisted Telephone Interviewing: Part I -- Introduction and Impact on Cost and Timeliness of Survey Data," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 93-115.
- ²⁷ Palit, Charles and Harry Sharp, "Microcomputer-Assisted Telephone Interviewing," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 169-189.

- ²⁸ Pallett, D. S. (editor), Automation of Data Capture for the Census in the Year 2000. Final report to the U.S. Bureau of the Census from the National Bureau of Standards under the interagency agreement for fiscal year 1987.
- ²⁹ Ponikowski, Chester, and Sue Meily, "Use of Touchtone Recognition Technology in Establishment Survey Data Collection." Presented at the First Annual Field Technologies Conference, St. Petersburg, Florida, 1988.
- ³⁰ Schuman, Howard and Stanley Presser. Questions and Answers in Attitude Surveys -- Experiments on Question Form, Wording, and Context. Orlando, FL: Academic Press, Inc., 1981.
- ³¹ Shanks, J. Merrill, "The Current Status of Computer-Assisted Telephone Interviewing," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 119-142.
- ³² Sharp, Harry, and Charles Palit, "Sample Administration with CATI: The Wisconsin Survey Research," Journal of Official Statistics, Vol. 4, No. 4, 1988, pp. 401-413.
- ³³ Sudman, Seymour, "Survey Research and Technological Change," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 217-230.
- ³⁴ Swann, T. C., "Electronic Data Collection in the Petroleum Supply Reporting System." Presented at the American Statistical Association Committee on Energy Statistics, April 28-29, 1988.
- ³⁵ Tortora, Robert D., "CATI in an Agricultural Statistical Agency," Journal of Official Statistics, Vol. 1, No. 2, 1986, pp. 301-314.
- ³⁶ Tucker, Clyde, "Interviewer Effects in Telephone Surveys," Public Opinion Quarterly, Vol. 47, 1984, pp. 84-95.
- ³⁷ U.S. Department of Agriculture, A Comparison of CATI and NonCATI on a Nebraska Hog Survey. Statistical Reporting Service Staff Report No. 89, by Richard Coulter. Washington, D.C.: Statistical Reporting Service, May, 1985.
- ³⁸ U.S. Department of Agriculture, Computer Assisted Telephone Interviewing on the Cattle Multiple Frame Survey. Statistical Reporting Service Staff Report No. 82, by Carol C. House. Washington, D.C.: Statistical Reporting Service, October, 1984.
- ³⁹ U.S. Department of Commerce, The Impact of High Technology on Data Collection. CATI Research Report N. GEN-1, by William L. Nicholls II. Washington, D.C.: U.S. Bureau of the Census, February, 1989.

- ⁴⁰ U.S. Department of Commerce, Proceedings of the Bureau of the Census First Annual Research Conference, Washington, D.C.: Bureau of the Census, 1985. Cost and Error Modeling for Large-Scale Surveys, by Robert M. Groves and James M. Lepkowski, pp. 330-357.
- ⁴¹ U.S. Department of Commerce, Proceedings of the Bureau of the Census Third Annual Research Conference, Washington, D.C.: Bureau of the Census, 1987. Use of Historical Data in a Current Interview Situation, by Bradley V. Pafford and Dick Coulter, pp. 281-298.
- ⁴² U.S. Department of Commerce, Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, D.C.: Bureau of the Census, 1988. "Discussion" of papers by Rothschild and Wilson and by Sebestik et al., by William L. Nicholls II, pp. 340-342.
- ⁴³ U.S. Department of Commerce, Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, D.C.: Bureau of the Census, 1988. Initial Experiences with CAPI, by Jutta Sebestik, Harvey Zelon, Dale DeWitt, James M. O'Reilly and Kevin McGowan, pp. 357-365.
- ⁴⁴ U.S. Department of Commerce, Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, D.C.: Bureau of the Census, 1988. Nationwide Food Consumption Survey Using Laptop Computers, by Beth B. Rothschild and Lucy B. Wilson, pp. 347-356.
- ⁴⁵ U.S. Department of Commerce, Proceedings of the Bureau of the Census 1990 Annual Research Conference, Washington, D.C.: Bureau of the Census, 1990. Building Predictive Models of CAPI Acceptance in a Field Interviewing Staff, by Mick Couper, Robert M. Groves and Curtis A. Jacobs, forthcoming.
- ⁴⁶ U.S. Department of Commerce, Proceedings of the Bureau of the Census 1990 Annual Research Conference, Washington, D.C.: Bureau of the Census, 1990. The Impact of Microcomputers on Survey Processing at the Netherlands Central Bureau of Statistics, by Wouter J. Keller and Jelke G. Bethlehem, forthcoming.
- ⁴⁷ U.S. Department of Commerce, Proceedings of the Bureau of the Census 1990 Annual Research Conference, Washington, D.C.: Bureau of the Census, 1990. Improving Data Quality in National Surveys: Experience with Computer-assisted Methods in the National Post-secondary Student Aid Studies, by James E. Smith and Carmen Vincent, forthcoming.
- ⁴⁸ U.S. Department of Labor, Voice Recognition and Voice Response Applications for Data Collection in a Federal/State Establishment Survey. By Richard Clayton and Debbie Winter. Washington, D.C.: U.S. Bureau of Labor Statistics, 1989.

⁴⁹ vanBastelaer, Alois, Frans Kessemakers and Dirk Sikkel, "Data Collection with Hand-Held Computers: Contributions to Questionnaire Design," Journal of Official Statistics, Vol. 4, No. 2, 1988, pp. 141-154.

⁵⁰ Weeks, Michael F., "Call Scheduling with CATI: Current Capabilities and Methods," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

⁵¹ Werking, George, Alan Tupek and Richard Clayton, "CATI and Touchtone Self-Response Applications for Establishment Surveys," Journal of Official Statistics, Vol. 4, No. 4, 1988, pp. 349-362.

DISCUSSION

William L. Nicholls II
U. S. Bureau of the Census

Marc Tosiano's paper has a didactic purpose. He presents basic information about CATI and related topics as background for a more technical paper on computer assisted survey information collection (CASIC) to follow. Since his paper is primarily a condensation of summary articles on CATI and CAPI previously prepared by others, it contains much that is familiar and little that is original. Rather than add another layer of commentary to this well worked material, I will use the discussant's time to counterpose the tone of technological and methodological optimism which seems to characterize many papers of this conference with some historical reality. This also will be familiar material to some readers, since it is based on the same sources as Tosiano's paper.

CATI and its associated technologies provide many opportunities to improve the timeliness and quality of survey data, often at the same or lower cost per case (Catlin and Ingram, 1988; Nicholls and Groves 1986; Groves and Nicholls 1986). But those increasingly documented benefits have not necessarily prompted Federal data collection agencies to implement CATI expeditiously in their major surveys or in ways that optimize those benefits.

The first CATI survey was conducted by Chilton Research in 1971; and by 1980 CATI was in widespread use in commercial market research and in university survey research (Nicholls, 1988). But even those Federal agencies moving most quickly, such as NASS, will not fully implement CATI in the their major continuing surveys before 1992. That will be 21 years, or a full generation, after CATI was invented. For the Census Bureau's major household surveys, such as the Current Population Survey and the National Crime Survey, the earliest conceivable date for full CATI (and CAPI) implementation is 1994, but slippage, say to 1996, seems increasingly likely under current budgetary constraints. That would represent a quarter century after the first CATI survey in the private sector. Federal agencies have introduced CATI more quickly into new and infrequently conducted surveys. But why has it taken so long to implement CATI for major, continuing Federal surveys?

There are many reasons. In the early 1970s, according to Dillman and Tarnai (1988), the managers of most Federal surveys regarded the telephone interview as a generally inferior data collection method and were reluctant to try it. Where a readiness for change was present, the technology often was lacking. CATI software was initially designed for market research and was not adequate for many government applications until enhanced by university organizations with government support in the late 1970s.

When U.S. government agencies began active internal development of CATI, around 1980, they often started with research programs to assess its effects on costs, data quality, and estimates. This research still continues, although both the Census Bureau and Statistics Canada produced major summaries of results in the late 1980s (U.S. Census Bureau, 1987; and Catlin and Ingram, 1988). The familiar delays of government planning, budgeting, and procurement also undoubtedly played a role in delaying CATI implementation.

CATI's extended incubation period in government also may be partly explained by its initial association with two related methodologies, random digit dialing (RDD) and centralized telephone interviewing, which also are topics of this session. Together, RDD, centralized telephone interviewing, and CATI are sometimes described as "modern telephone methods." Their joint evolution was described by Groves and Kahn in their influential 1979 volume Surveys by Telephone as one of the major developments in the history of survey methods, ranking with area probability sampling and the use of computers for survey analysis. By 1980, Berry and O'Rourke (1988), among others, have noted that modern telephone methods (RDD, centralized, and with CATI) had become the dominant survey methods in U.S. commercial market research and in university survey research centers. Government agencies were the exception.

"Modern telephone methods" did not transfer readily to government data collection as a package. This is most apparent for random digit dialing, whose potential to reduce survey costs attracted major interest among government statisticians (Biemer et al., 1985). The National Center for Health Statistics, the Census Bureau, and Statistics Canada all began their investigations of modern telephone methods with years of careful testing of random digit dialing (Marquis and Blass, 1985). But, as Drew, Choudhry, and Hunter (1988) have observed, RDD sampling methods are used in few government surveys conducted in the U.S. or elsewhere in the world. The omission of nontelephone households (about 7 percent of the U.S. total) and the typically higher refusal rates of cold contact telephone interviews have presented major barriers to the use of RDD in many or most government survey applications.

Random digit dialing remains a valuable sampling method for populations with high telephone subscribership (such as Canada and Sweden) and for surveys which can tolerate its coverage and nonresponse problems. For some governmental statistical agencies, however, the early emphasis on RDD proved a diversion from what now appear to be more fruitful uses of CATI. Only when RDD was ruled out as a sampling method for most U.S. government household surveys, which at the Census Bureau occurred around 1986, could plans to implement CATI in single-frame, mixed mode designs proceed. The somewhat faster adoption of CATI by establishment and agricultural surveys may be partly attributable to their traditional reliance on list frame samples. A change to RDD was not an issue.

The second major element of modern telephone methods which has not translated easily to government data collection is centralized telephone interviewing. In U.S. university and commercial market research, the shift from "dispersed" local interviewers making calls from their own homes to "centralized" telephone interviewers calling from large national or regional offices was largely completed by the late 1970s. Government household surveys are one of the few major users of dispersed telephone interviewing persisting into the 1980s.

Mr. Tosiano's paper has reviewed the ways in which centralized telephone interviewing and CATI can be mutually supporting methodologies. Computer-assistance is easier to arrange for centralized interviewers who share the same hardware, programs, sample, and technical staff. At the same time, CATI encourages centralized interviewing to gain these efficiencies and to benefit from such large-staff CATI features as automatic call scheduling, online supervision, and field report generation. Centralization contributes to standardized field procedures and interviewing quality control through easier recruitment, training, and supervision of interviewers, while CATI contributes to these same goals through tailored question wordings, computer controlled branching, and online editing. Supervisory audio-visual monitoring of interviewer performance, currently feasible only with centralized CATI interviewers, provides feedback ensuring that CATI quality enhancement features are appropriately used and that interviewers deviating from performance standards are identified and retrained when necessary.

"Centralization" has a different meaning for government establishment surveys than for government household surveys. Because the establishment surveys typically began with mailed questionnaire methods, later supplemented with telephone prompting and interviews, they generally are conducted from offices. The choice typically is between national, regional, and state offices. The introduction of CATI strengthens the arguments for greater centralization. The Census Bureau's Business Division is perhaps unique among Federal agencies in withdrawing its Retail and Wholesale Trade Surveys from a set of regional offices to centralize them in a national site before placing them on CATI. More commonly, existing organizational arrangements, Federal-State agreements, and formal or informal commitments to employees have resulted in continuation of state-based offices averaging about 10 interviewing stations per state but ranging from 2 to perhaps 30 stations (Nicholls 1988). In national private sector survey and market research organizations, CATI installations more typically reach 45-100 stations.

The introduction of CATI into mixed-mode personal-telephone household surveys presents even greater organizational problems. This is illustrated by the Census Bureau's plans to phase CATI into the Current Population Survey (CPS) and the National Crime Survey

(NCS). Both surveys have a rotating panel design. The first visit to each sample address is by personal visit to identify ineligible housing units and to encourage household participation. The fifth CPS and NCS visits also are in person to re-establish personal contact with the household part way through the sequence of interviews. Other interviews are by telephone when possible and acceptable to the respondent and by personal visit otherwise. The same local interviewers traditionally conduct both the personal and telephone visit interviews, placing the telephone interview calls from their own homes.

When CATI is introduced into these surveys, no change is made in the initial visits to each sample address. These remain personal visit interviews since comparable response and panel retention rates have not been attainable with cold contact telephone interviews (Marquis and Blass, 1985). CATI replaces dispersed telephone interviews from the local interviewers' homes in the second and later visits of the panel design. This field design has several potential benefits: (1) reduced field costs; (2) reduced interviewer recruitment problems in tight labor markets; and (3) possibly improved survey estimates. Nevertheless, the transition poses a number of design and organizational problems which require time and effort to resolve.

The first is developing appropriate methods for rapid but controlled transfer of individual case records between personal visit and CATI interview modes. When the first-visit personal interview is complete, household enumeration data and field records must be data entered into computer files for second and later interviews by CATI. Case records also move from CATI to the local interviewers for CPS and NCS fifth visits and for personal followup of households unreachable by CATI.

The second transition problem is the temporarily reduced efficiency of the sample designs. Both the CPS and NCS employ cluster samples chosen initially to minimize costs for interviewing assignments containing both personal visit and dispersed telephone interviews. When the dispersed telephone interviews are removed to CATI, the remaining personal visit cases may no longer constitute acceptable or efficient field assignments. Since the CPS and NCS samples are based on the decennial census, they are efficiently revised only once a decade.

The third problem in moving dispersed telephone interviews to CATI is the need to reduce the field staff while increasing the CATI staff. For the CPS, the Census Bureau's largest current survey, the transition will be based initially on field interviewer attrition and has been constrained by the rate at which attrition occurs.

The fourth and final transition problem is finding a sufficient volume of work for the CATI interviewers. The CPS

conducts its interviews in the third week of each month and the NCS in the first week with some carryover into the second. Centralized CATI interviewing is restricted to even fewer days per month to permit field followup of cases unreachable by telephone. These two surveys will provide the CATI staff with relatively few days of employment per month.

Of the four transition problems, only the first derives from the CATI technology. Case transfers between dispersed local interviewers and centralized CATI interviewers are complicated by the move between paper-and-pencil records and computer files. The problems of field sampling efficiency, field staff phase-down, and insufficient work at the CATI facilities arise from the centralization of previously dispersed interviews. They would be the same whether the central facility used CATI or paper-and-pencil methods.

The most difficult problems of implementing CATI in government agencies appear to derive from the organizational issues CATI typically raises about centralized vs. decentralized interviewing.

References

Biemer, P., D.W. Chapman, and C. Alexander, "Some Research Issues in Random-Digit Dialing Sampling and Estimation," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 71-86.

Berry, S. H. and D. O'Rourke, "Administrative Designs for Centralized Telephone Survey Centers: Implications of the Transfer to CATI," in R. M. Groves et al. (editors) Telephone Survey Methodology, New York, Wiley Press, 1988, pp. 457-474.

Catlin, G. and S. Ingram, "The Effect of CATI on Cost and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing," in R. M. Groves et al. (editors) Telephone Survey Methodology, New York, Wiley Press, 1988, pp. 437-452.

Dillman, D. A. and J. Tarnai, "Administrative Issues in Mixed Mode Surveys," in R. M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988, pp. 509-528.

Drew, J. D., G. H. Choudhry, and L. A. Hunter, "Nonresponse Issues in Government Telephone Surveys," in R. M. Groves et al. (editors) Telephone Survey Methodology, New York, Wiley Press, 1988, pp. 233-246.

Groves, R. M. and R. L. Kahn, Surveys by Telephone, New York, Academic Press, 1979.

Groves, R. M. and W. L. Nicholls II, "The Status of Computer-Assisted Telephone Interviewing: Part II -- Data Quality Issues," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 117-134.

Marquis, K. and R. Blass, "Nonsampling Error Considerations in the Design and Operation of Telephone Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 301-329.

Nicholls II, W. L., "The Impact of High Technology on Survey Data Collection," CATI Research Report GEN-1, U.S. Department of Commerce, Bureau of the Census, February 1989.

Nicholls II, W.L. and Groves, R.M., "The Status of Computer-Assisted Telephone Interviewing: Part I -- Introduction and Impact on Cost and Timeliness of Survey Data," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 93-115.

U.S. Census Bureau, Evaluation of CATI Data Quality and Costs in the Current Population Survey, CATI Research Report No. CPS-2 of the Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee, September 1988.

DISCUSSION

James T. Massey
National Center for Health Statistics

The paper by Leyla Mohadjer and David Morganstein enumerates and provides a brief overview of almost all of the key methodological issues related to telephone surveys. The concept of total survey design mentioned in the first section of the paper is an excellent way to compare and summarize the advantages and disadvantages of telephone surveys versus other modes of data collection. The total survey design concept was never fully developed to compare the different modes of data collection. Most of this paper focused on the operational and sample design efficiencies of telephone surveys to improve data quality.

The advantages and disadvantages of telephone surveys given by Mohadjer and Morganstein are listed below along with several additional ones:

Advantages of Telephone Surveys

- Lower cost
- Better quality control and supervision of interviewers
- Better access to some hard to reach persons
- Smaller design effects
- Smaller interviewer effects
- Cost effective method to sample rare population (use of Donnelly tape with characteristics of persons in prefix area)
- Use of CATI to control flow of sample, interview, edits, and processing
- Local area surveys from central location
- Better use of bilingual interviewers

Disadvantages of Telephone Surveys

- Lack of visual aids
- No group interviews

- Noncoverage of persons without telephones
- Lower response rates
- Cost of dual frame surveys
- Cost of CATI (relative to other telephone surveys)

Now I would like to turn my attention to where we are in the development and use of telephone surveys and some areas that still need research.

I see six reasons for the emergence of telephone surveys:

- 1) Better coverage
- 2) Development of CATI which lead to development of CAPI
- 3) Development of better RDD methods
- 4) Higher costs to fact-to-face surveys
- 5) Slow death of the myth of the length of a telephone interview
- 6) Recognition of data quality equal to face-to-face surveys

Considerable progress has been made over the past 15 years in almost every aspect of telephone surveys including data quality. There are, however, several areas where progress has been limited and more research is needed. These are listed below.

- 1) Techniques to improve response rates: While response rates have improved, there is still much that could be done to adapt procedures in the face-to-face surveys to telephone surveys. I just reviewed a paper that used several inducement techniques to dramatically improve telephone survey response rates in another country.
- 2) Validation of data collected by telephone versus face-to-face surveys: Most comparative studies have assumed that higher levels of reporting is better. For some types of data this assumption is questionable and additional statistical validation studies are needed.
- 3) Research on the collection of sensitive data and other specific types of information: We should take full advantage of one of the key features of telephone interviewing, the autonomy and anonymity of the interview. Some research has been done that showed sensitive data and questions have socially desirable

responses are obtained better over the telephone. There is some recent unpublished data that indicates that smoking habits, crimes, and unemployment may have higher reporting over the telephone. These results should be published and validated.

- 4) Research on difficult questions and questions with multiple response: Questions requiring flashcards and scaled responses are still more problematic over the telephone. CATI does offer a way to randomize the order of responses.
- 5) Research on better and cheaper ways to correct for noncoverage.

Finally, I would like to make two other observations. In 1984 when the OMB Working Paper 12 on Telephone Data Collection was published, telephone surveys were primarily used in the Federal government to conduct follow-up surveys and follow-up interviews. Most initial contact surveys by telephones used list frames. This is still the case today for almost all of the large government surveys, although greater use of telephone interviewing is being made.

For those of you who are new to the study of telephone survey, I recommend you start with the book Telephone Survey Methodology. It has several state-of-the-art review papers and has an extensive bibliography. The paper Owen Thornberry and I wrote contains as many reference tables on telephone coverage as Bob Groves would allow us to include. I hope many of you will continue to conduct research on telephone surveys and extend our knowledge of this very valuable data collection method.

